Le beau noiseur

Du bruit pour révéler le signal

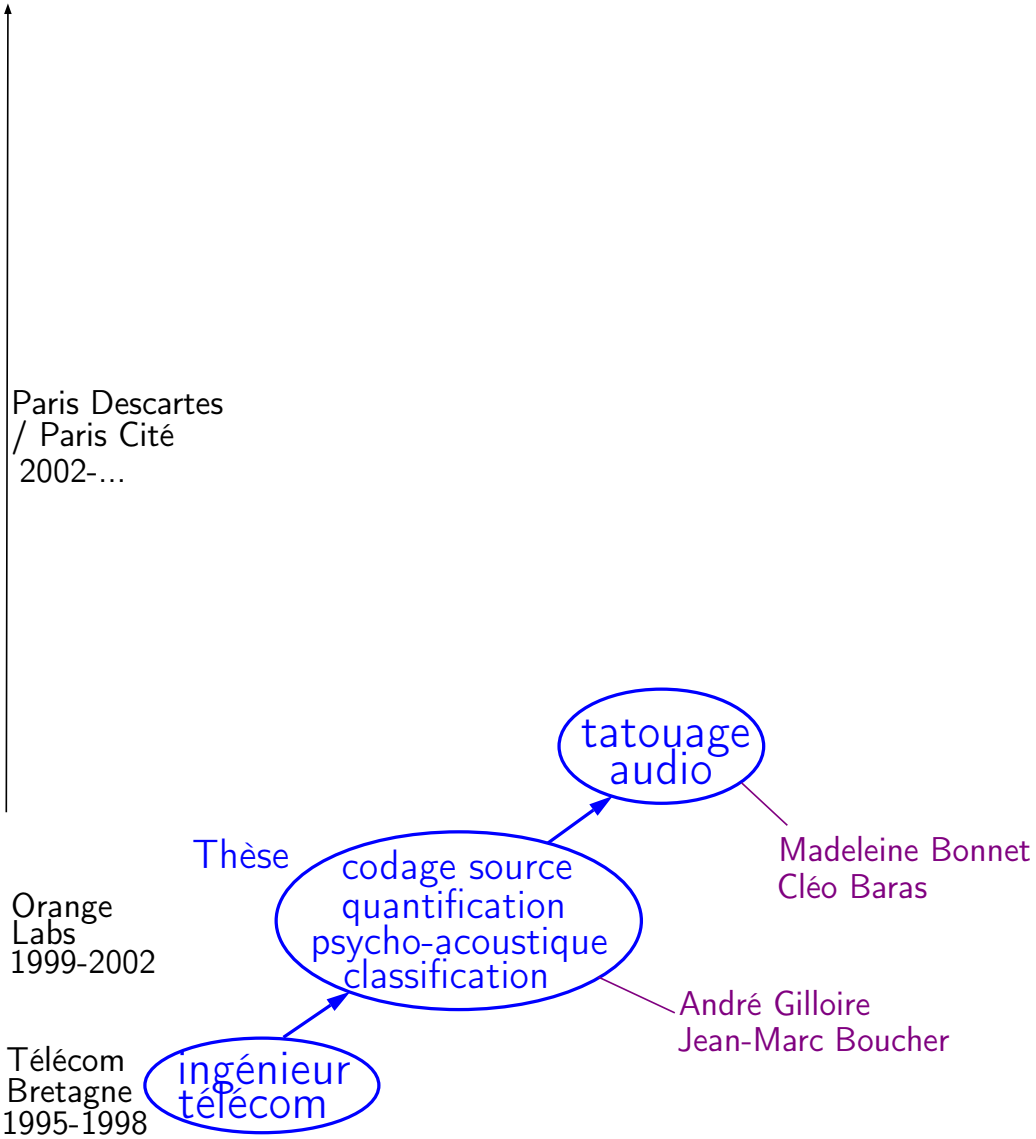
Synthèse des travaux de recherche
présentée pour la candidature
à l'habilitation à diriger des recherches

Gaël Mahé

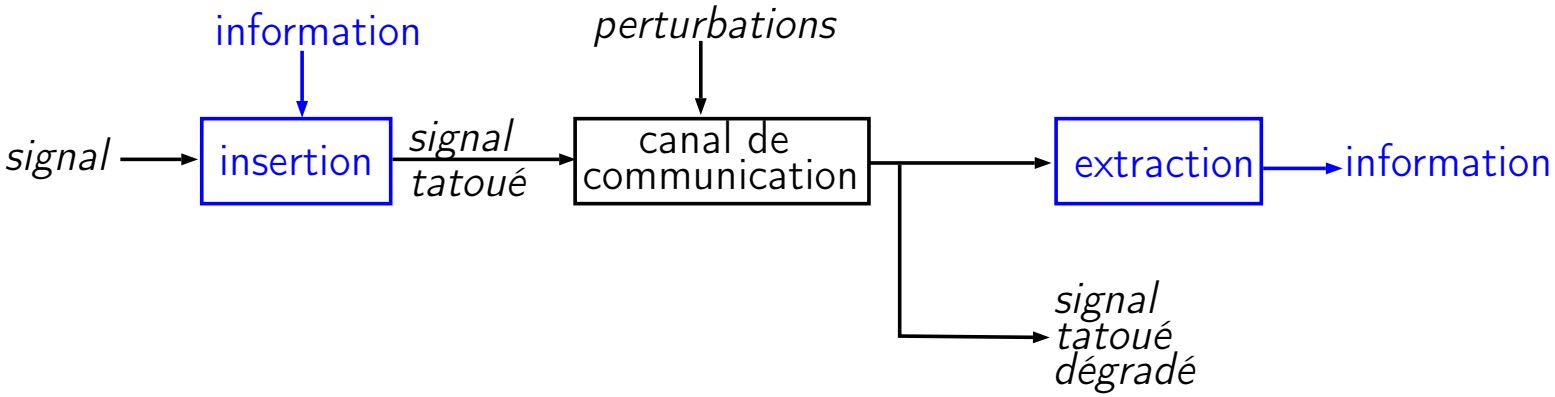
Université Paris Cité / Faculté des Sciences / UFR math-info / LIPADE

9 janvier 2023

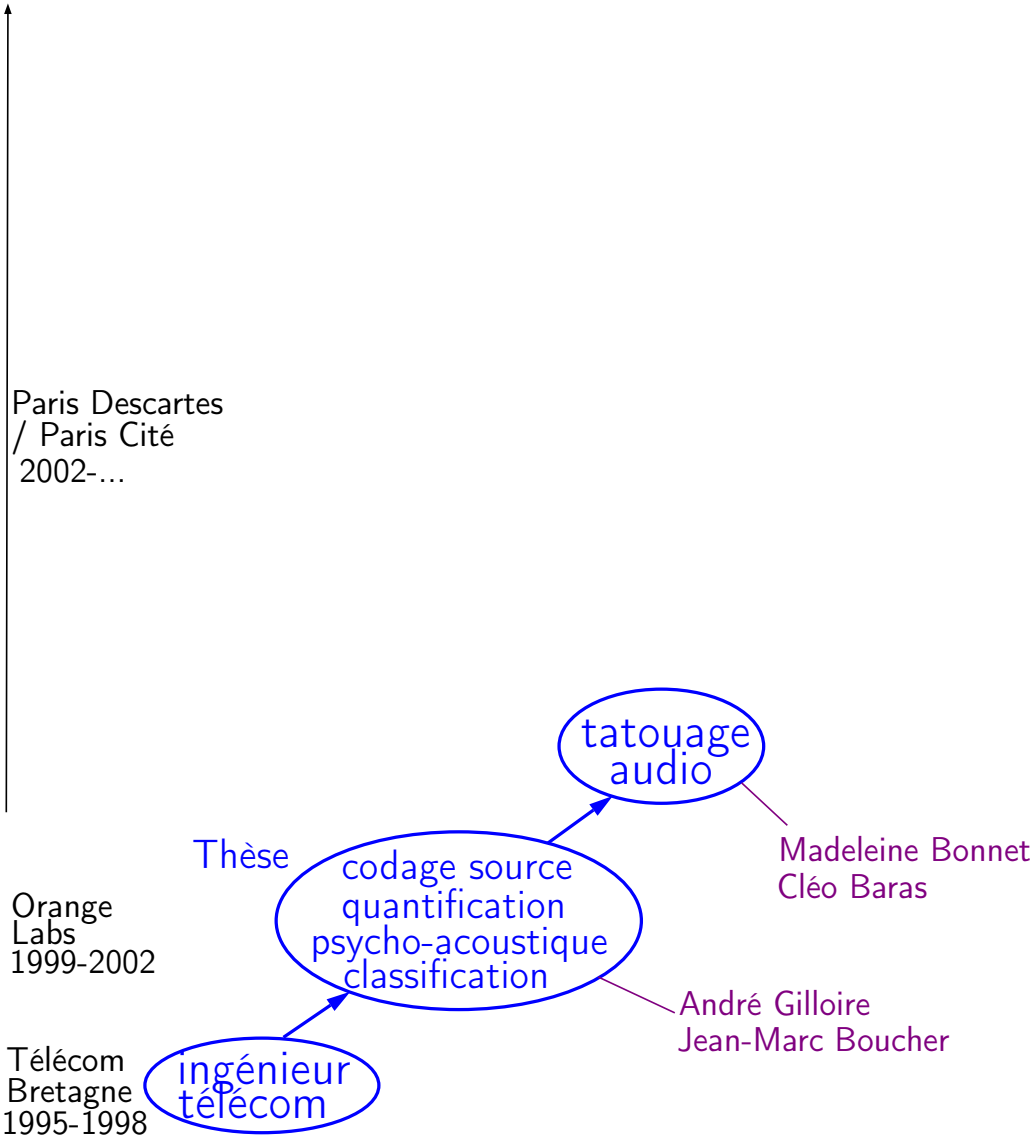
Introduction



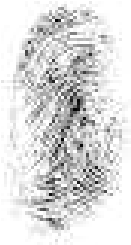
Tatouage audio



Introduction



Introduction



du bruit pour révéler

du bruit pour témoigner

du bruit pour doper

du bruit pour informer

rugby : analyse forces et mouvement

tatouage audio

Madeleine Bonnet
Cléo Baras

Thèse

codage source
quantification
psycho-acoustique
classification

André Gilloire
Jean-Marc Boucher

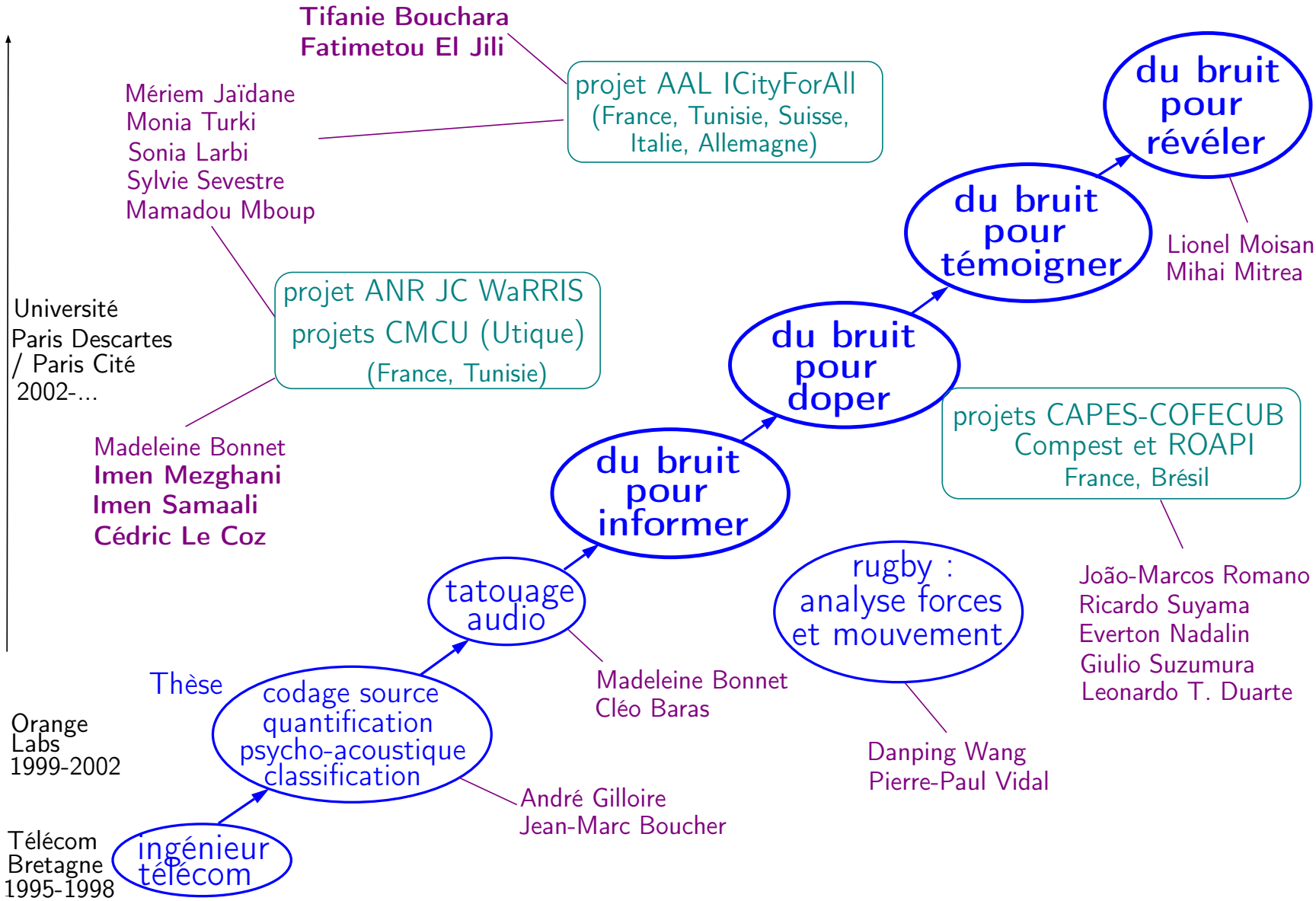
ingénieur télécom

Paris Descartes / Paris Cité
2002-...

Orange Labs
1999-2002

Télécom Bretagne
1995-1998

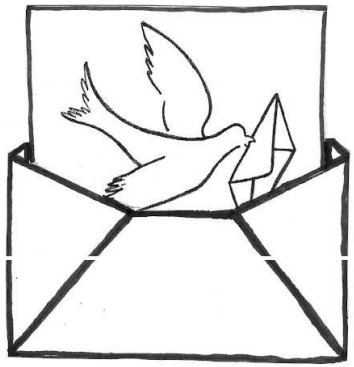
Introduction

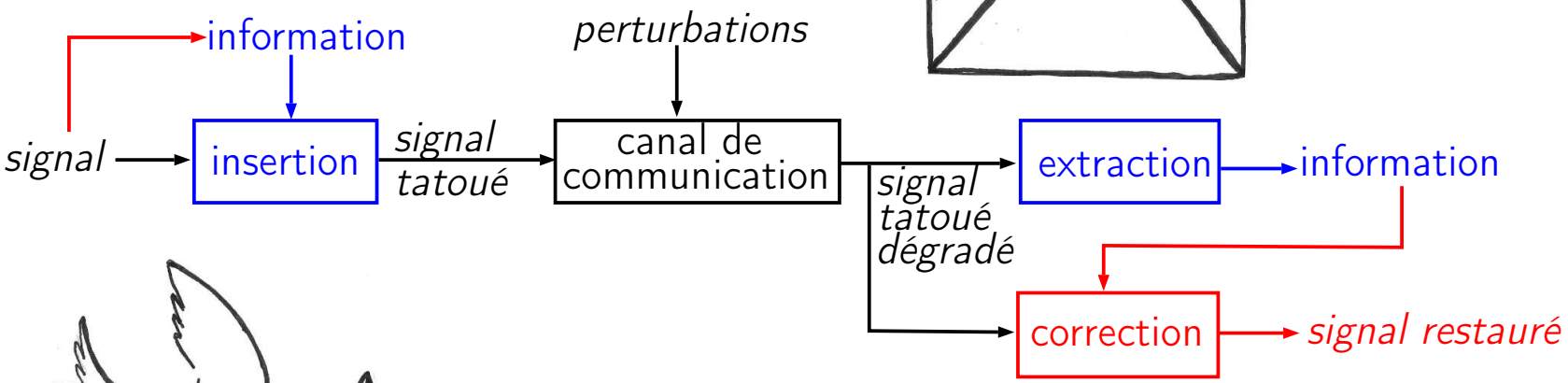


Plan

Plan

Du bruit pour informer : du tatouage au tatouage réflexif





Applications du tatouage réflexif

- Séparation de sources informée
- Restauration de tonales en codage audio
- Correction de pré-écho en codage audio

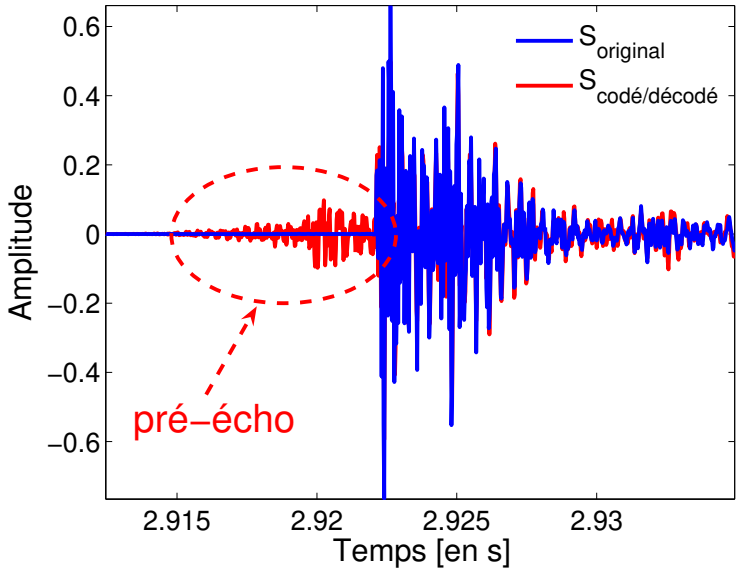
Applications du tatouage réflexif

- Séparation de sources informée
- Restauration de tonales en codage audio
- **Correction de pré-écho en codage audio**

Pré-écho dans les codecs MP3 et AAC

Quantification de la transformée

- bruit de quantification reformé en fréquence, mais uniforme en temps
- **pré-écho dans les attaques**



Castagnettes, codage MP3 à 48 kbit/s

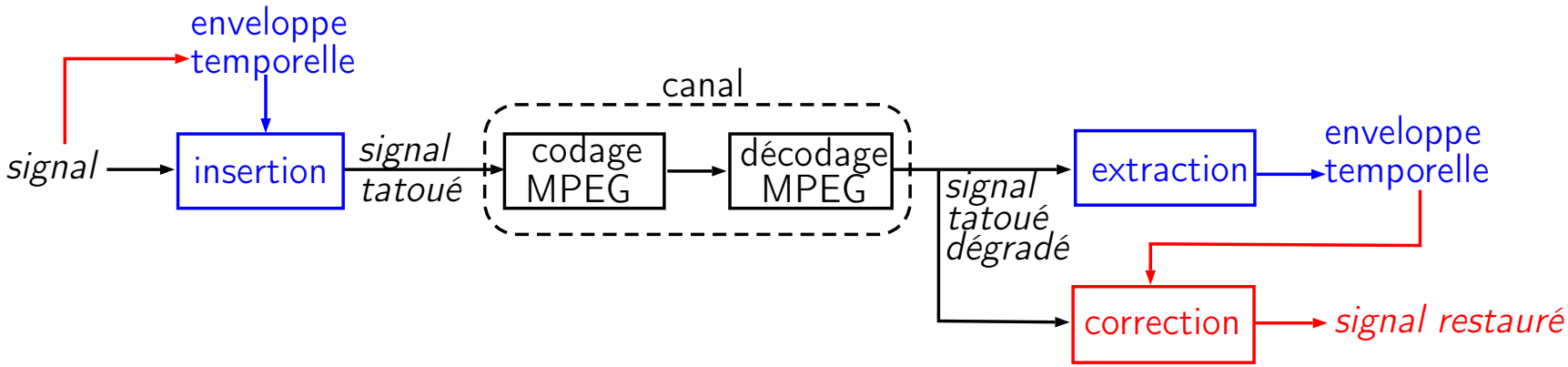
Comment éviter le pré-écho ?

- Inaudible si durée < 5ms $\rightarrow$ codage par fenêtres de longueur variable
  - MP3 : option Temporal Masking (TM)
- Quantification du résidu de prédiction linéaire du spectre
  - AAC : option Temporal Noise Shaping (TNS)
- Mais n'annule pas tous les pré-échos et augmente le débit

Correction de pré-écho par tatouage réflexif (1)

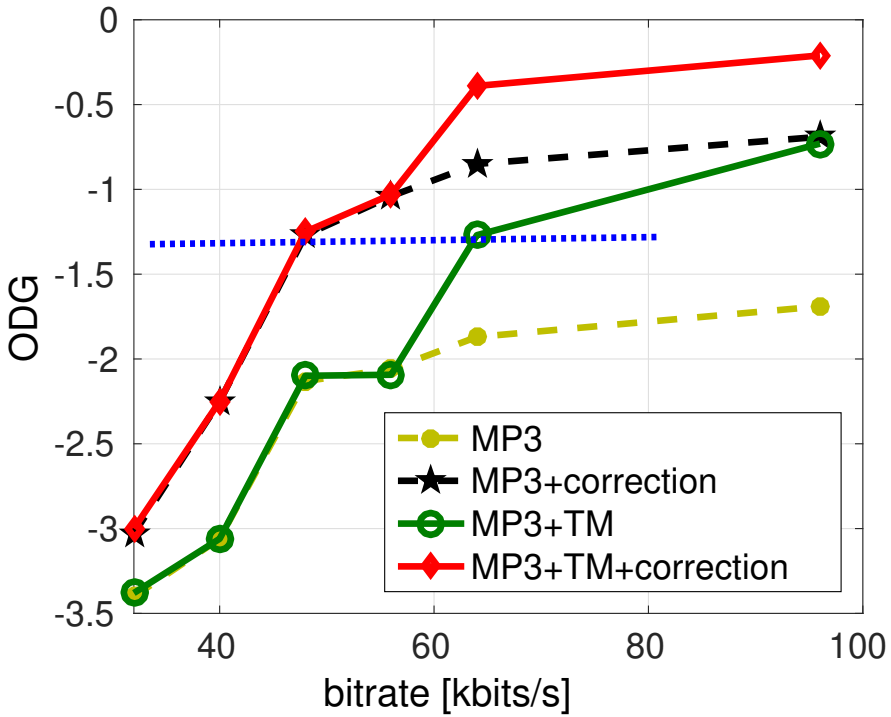
Principe : transmettre par tatouage l'enveloppe temporelle

I. Samaali, G. Mahé et M. Turki, "Watermark-aided pre-echo reduction in low bit-rate audio coding", J. of the Audio Engineering Society, 2012



Correction de pré-écho par tatouage réflexif (2)

Castagnettes, information auxiliaire tatouée à 50 bit/s, codage MP3



Original

Codé-décodé

Corrigé

Plan

Genèse

- Un tatouage par étalement de spectre stationnarise le signal hôte
  $\rightarrow$ amélioration performances d'un annuleur d'écho

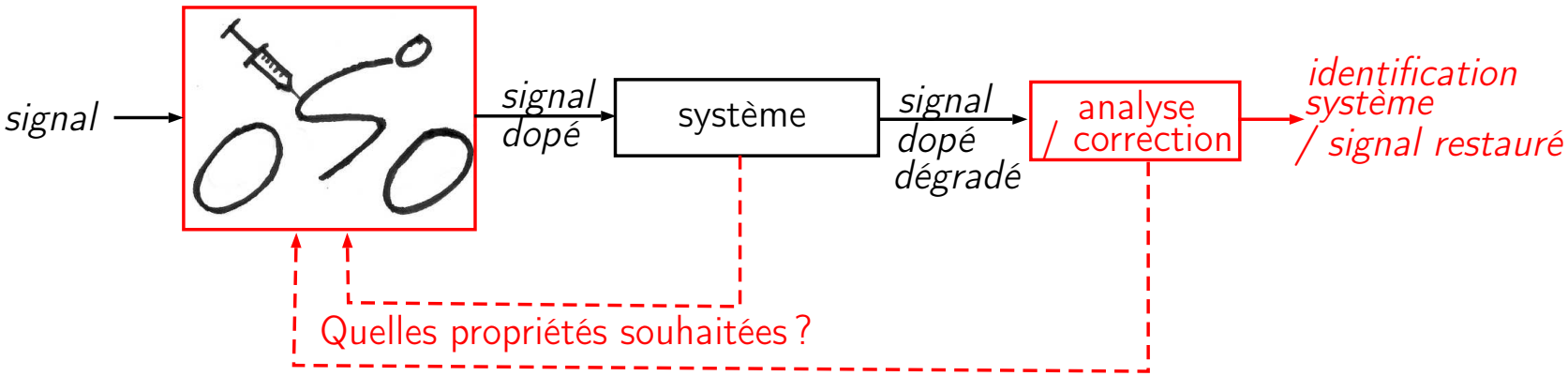
  S. Larbi et M. Jaïdane, Audio watermarking : A way to stationarize audio signals. IEEE Trans. Signal Processing, 2005
- Un bruit améliore les performances d'un annuleur d'écho stéréphonique en décorrélant les deux voies

  A. Gilloire et V. Turbin, Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers, ICASSP 1998

Idée

Concevoir un tatouage (ou un bruit) expressément pour adapter les propriétés d'un signal audio à un traitement ultérieur

Du bruit pour doper



Applications du dopage

- Parcimoniser et supergaussianniser
  $\rightarrow$ séparation de sources
- Gaussianniser
  $\rightarrow$ identification de systèmes non-linéaires
- Représenter le signal comme combinaison de codes correcteurs
  $\rightarrow$ débruitage
- Filtrer passe-bas l'histogramme d'un signal
  $\rightarrow$ application du théorème de quantification

Applications du dopage

- Parcimoniser et supergaussianniser
  $\rightarrow$ séparation de sources
- Gaussianniser
  $\rightarrow$ identification de systèmes non-linéaires
- Représenter le signal comme combinaison de codes correcteurs
  $\rightarrow$ débruitage
- **Filtrer passe-bas l'histogramme d'un signal**
  $\rightarrow$ **application du théorème de quantification**

Quantification et sous-quantification

Théorème de quantification (B. Widrow, 1956)

- x signal discret, quantifié avec un pas $q \rightarrow x_Q$
- Fonction caractéristique de x à support borné, fréquence max ν_{max}
- La densité de probabilité (ddp) de x peut être reconstruite parfaitement à partir de celle de x_Q si $1/q > 2 \times \nu_{max}$

Ici, sous-quantification

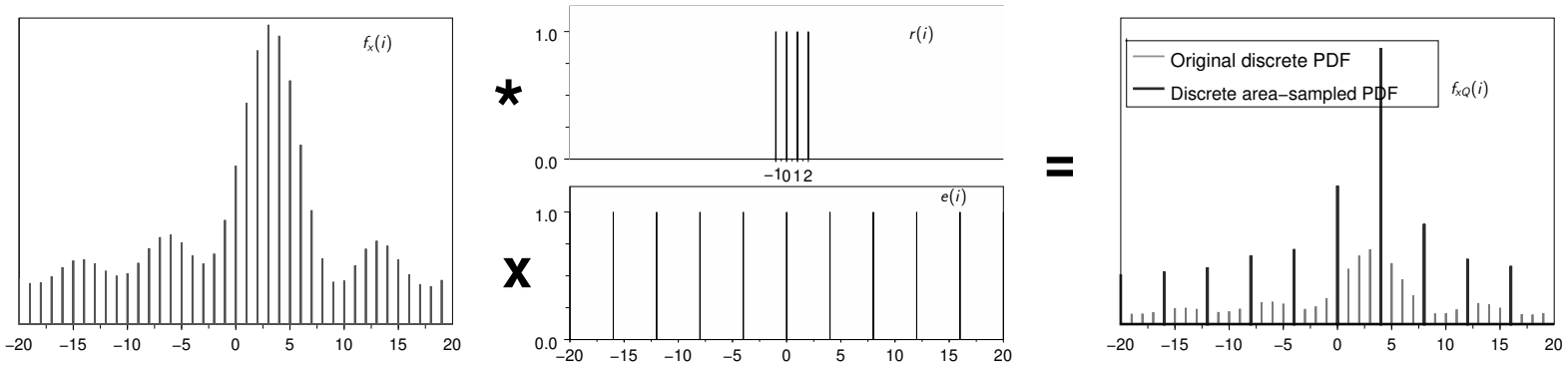
- Sous-quantifier d'un facteur K un signal x à valeurs entières

$$\forall\, x(t) \in \lfloor nK - K/2 + 1, nK + K/2 \rfloor, \quad x_Q(t) = nK$$

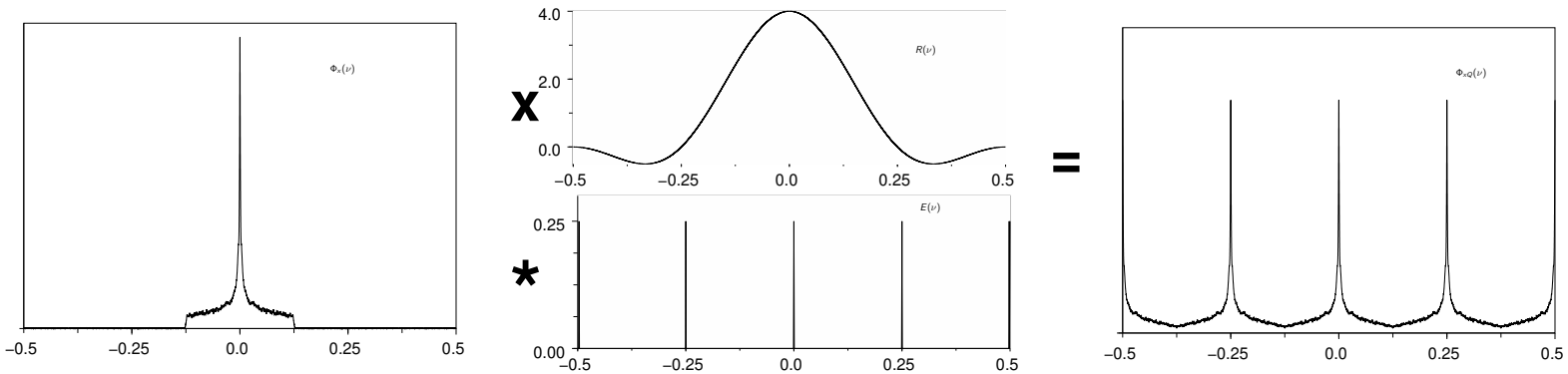
- ddp du signal sous-quantifié xQ :

$$\begin{aligned} f_{xQ}(nK) &= \sum_{i=\lfloor nK - \frac{K}{2} + 1 \rfloor}^{\lfloor nK + \frac{K}{2} \rfloor} f_x(i) \\ f_{xQ}(i \neq nK) &= 0 \end{aligned}$$

Sous-quantification = sous-échantillonnage de surface de la ddp



Équivalent dans le domaine fréquentiel :



Théorème de sous-quantification

- x signal quantifié, sous-quantifié d'un facteur K en x_Q
- **Si la fonction caractéristique de x est nulle $\forall\, |\nu| > \frac{1}{2K}$ sur $[-\frac{1}{2}; \frac{1}{2}]$ alors la ddp de x peut être déduite de celle de x_Q**
- Dans le domaine fréquentiel, cette restauration s'exprime :

$$\Phi_x(\nu) = \Phi_{xQ}(\nu) G(\nu) \tag{1}$$

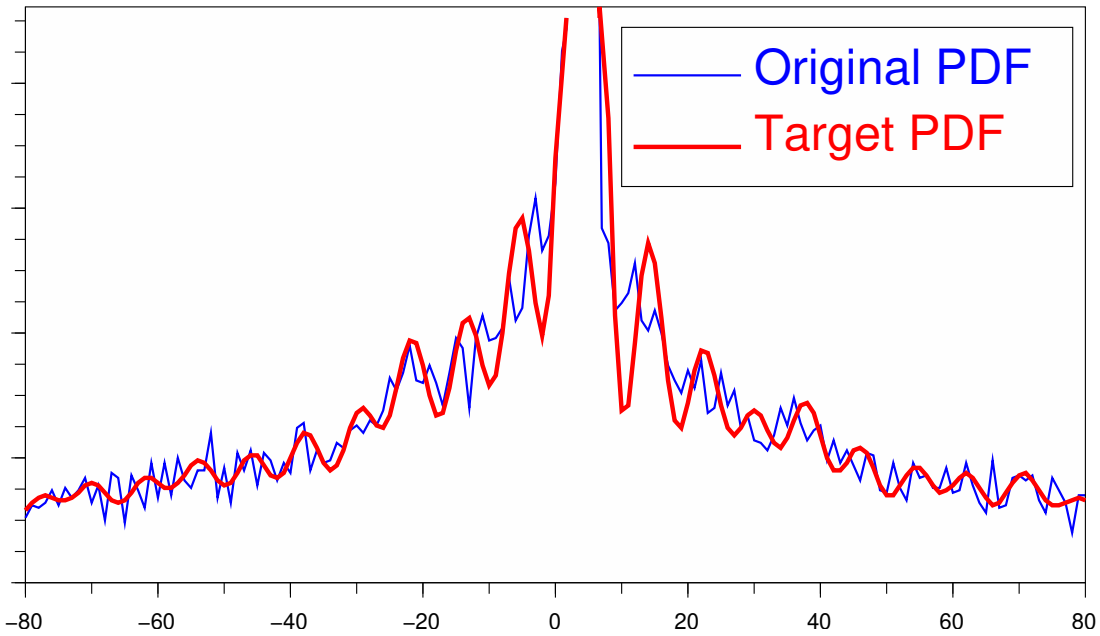
où Φ_{xQ} = fonction caractéristique de x_Q
et

$$G(\nu) = \begin{cases} 0 & \text{si} |\nu| > \frac{1}{2K} \\ 1/K & \text{si } \nu = 0 \\ \frac{sin(\pi\nu)}{sin(\pi K\nu)} \exp(j\pi\nu) & \text{si } \nu \neq 0,\ |\nu| < \frac{1}{2K} \text{ et } K \text{ pair} \\ \frac{sin(\pi\nu)}{sin(\pi K\nu)} & \text{si } \nu \neq 0,\ |\nu| < \frac{1}{2K} \text{ et } K \text{ impair} \end{cases} \tag{2}$$

- Valable aussi avec histogramme au lieu de ddp

Comment vérifier ces conditions ?

Filtrage passe-bas de la ddp de x, fréquence de coupure $\frac{1}{2K}$:



Contrôle perceptif du reformage d'histogramme

Objectif

- Soit f_{target} l'histogramme cible. Trouver une transfo $x \to z$ telle que :

$$\begin{cases} f_z \simeq f_{target} & (3) \\ w = z - x \text{ est inaudible} & (4) \end{cases}$$

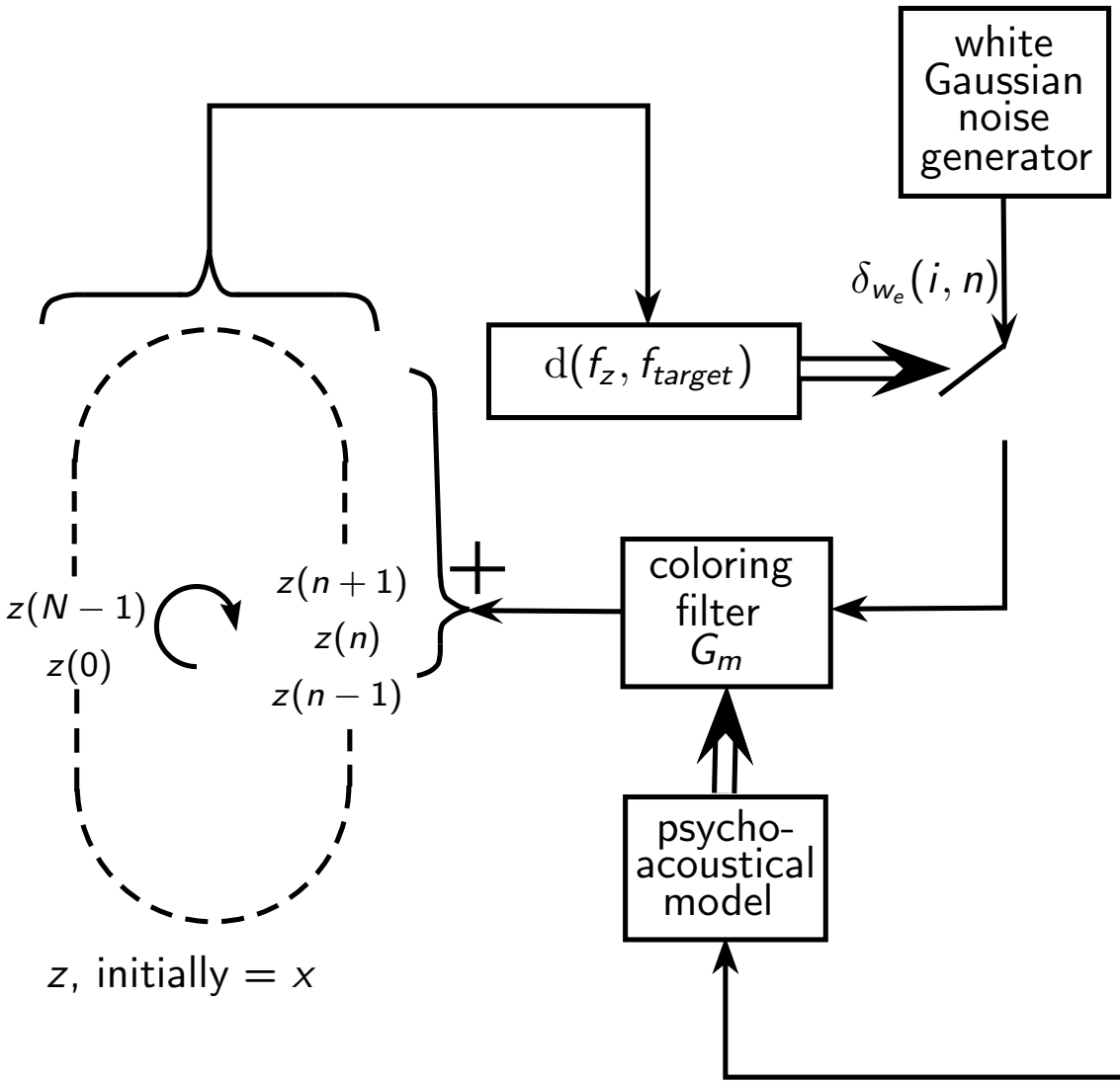
- En d'autres termes,

$$\begin{cases} \min \mathrm{d}(f_z, f_{target}) \quad \text{sous la contrainte :} & (5) \\ \gamma_w(m,\nu) < \gamma_{mask}(m,\nu) \quad \forall \,\text{trame}\, m,\, \text{fréquence}\, \nu & (6) \end{cases}$$

Difficulté

1. Optimisation d'un histogramme défini sur tout le signal
   vs contraintes locales (un seuil de masquage par trame)
2. Optimisation exprimée à partir de la représentation temporelle
   vs contraintes exprimées dans le domaine fréquentiel

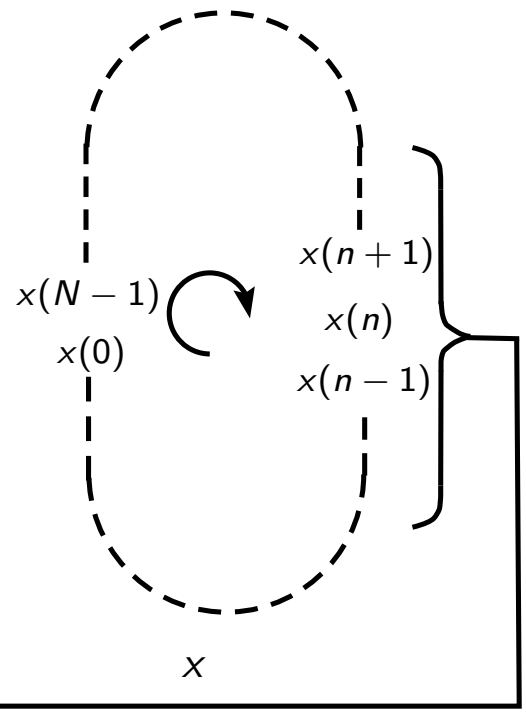
Algo de reformage d'histogramme perceptivement contrôlé



À la fin, $w = g_m * w_e$,
avec w_e bruit blanc

Contrainte $\gamma_w(\nu) < \gamma_{mask}(\nu)$

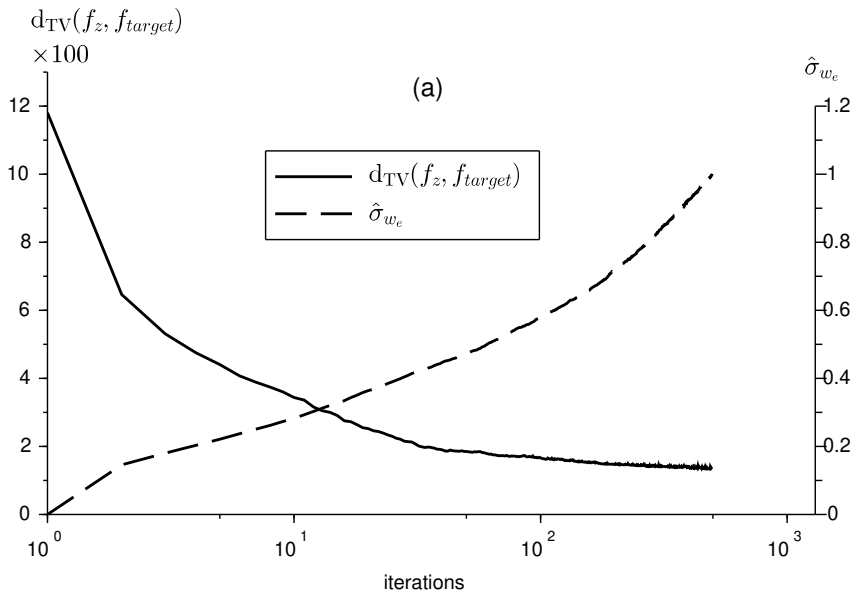
$$\Leftrightarrow \begin{cases} |G_m(\nu)|^2 = \gamma_{mask}(\nu) \\ \sigma_{w_e} < 1 \end{cases}$$



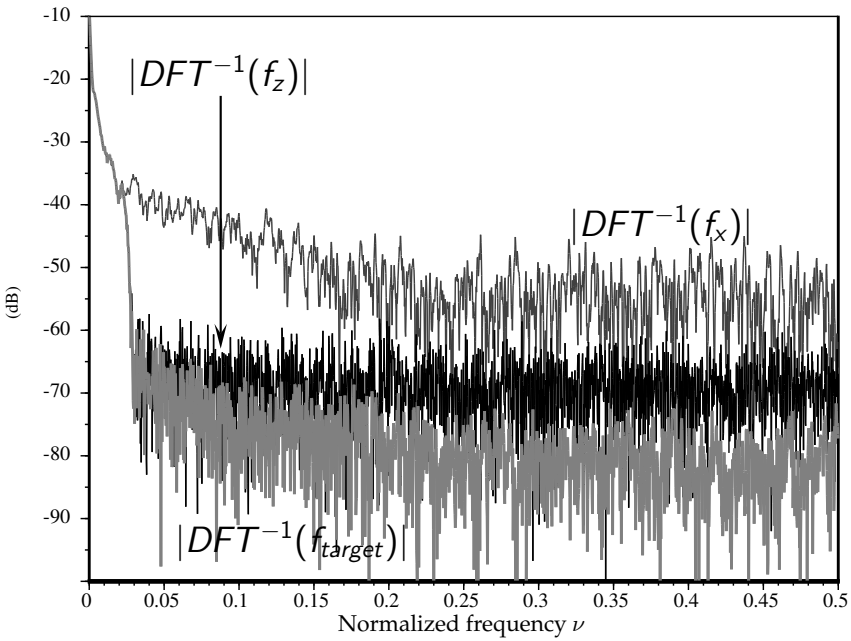
Résultats

- Signal de parole de 3 s, à 16 kHz
- Filtrage passe-bas de l'histogramme, fréquence de coupure 1/32
- Après 100 itérations, MOS (estimé par PESQ) entre 4 et 4,5

Évolution de $\mathrm{d_{TV}}(f_z, f_{target})$ et de la variance de w_e au fil des itérations



Après 100 itérations, fonctions caractéristiques



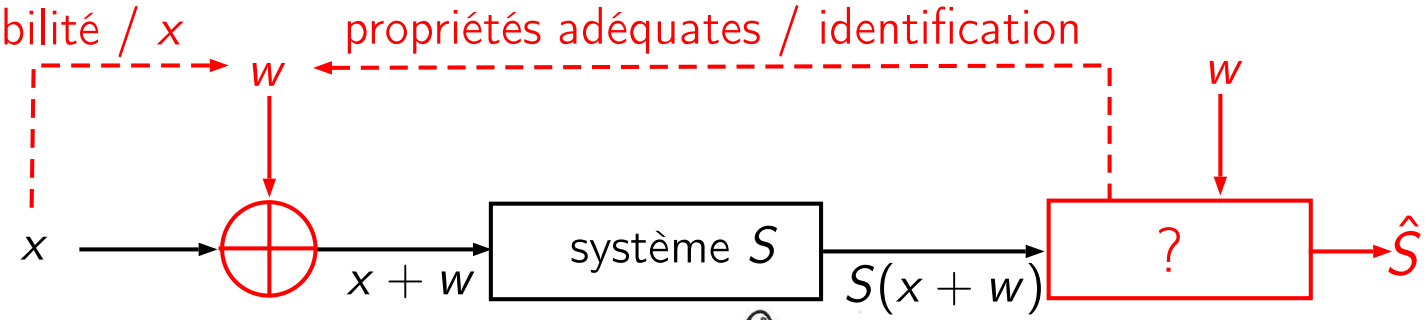
G. Mahé et M. Jaidane. Perceptually Controlled Reshaping of Sound Histograms. IEEE/ACM Trans. on Audio, Speech and Language Processing, 2018

Quel intérêt ?

- Intérêt théorique : **transposition des outils de l'échantillonnage** à la quantification
- Intérêt pratique : **ciselage de la ddp** ou de la fonction caractéristique
  $\rightarrow$ codage, tatouage
- Intérêt théorique et pratique :
  Extension du théorème de sous-quantification à la ddp jointe
  $f\left(x(n), x(n-1), \ldots, x(n-p)\right)$
  $\rightarrow$ **déquantification**

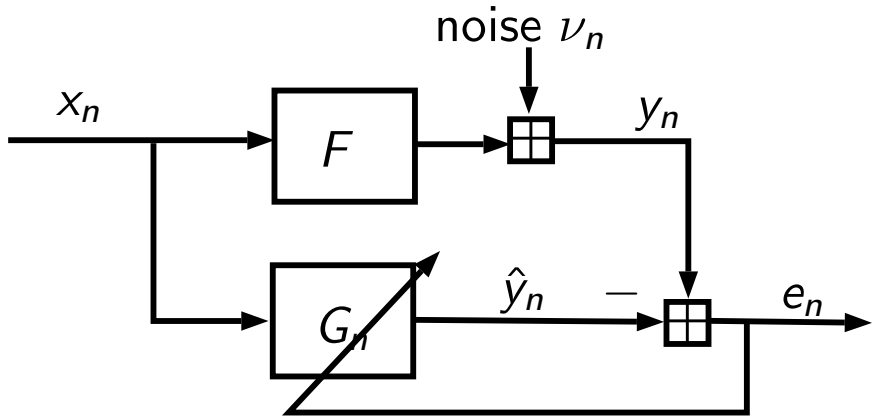
Plan

Le bruit témoin des altérations du signal



w s'imprègne des mêmes altérations que le signal hôte

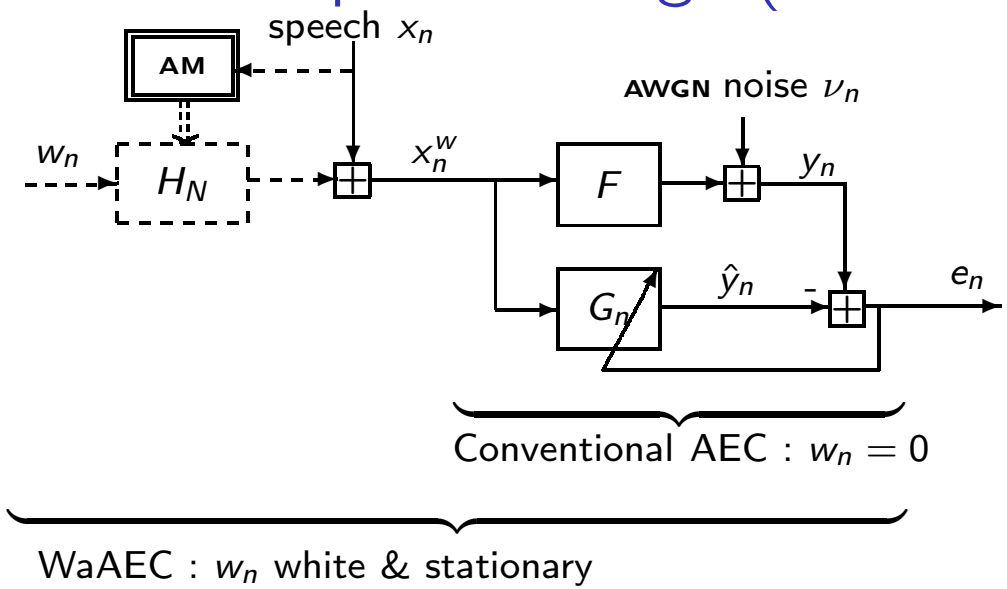
Identification de système linéaire (AEC)



Identification de F par filtrage adaptatif
ex. NLMS : $G_{n+1} = G_n + \frac{\mu}{\|X_n\|^2} e_n X_n$

- Vitesse de convergence dépend de la blancheur du signal
- Performances en régime permanent dépendent de la stationnarité
- Or le signal de parole n'est ni blanc ni stationnaire...

Annuleur d'écho assisté par tatouage (WaAEC)

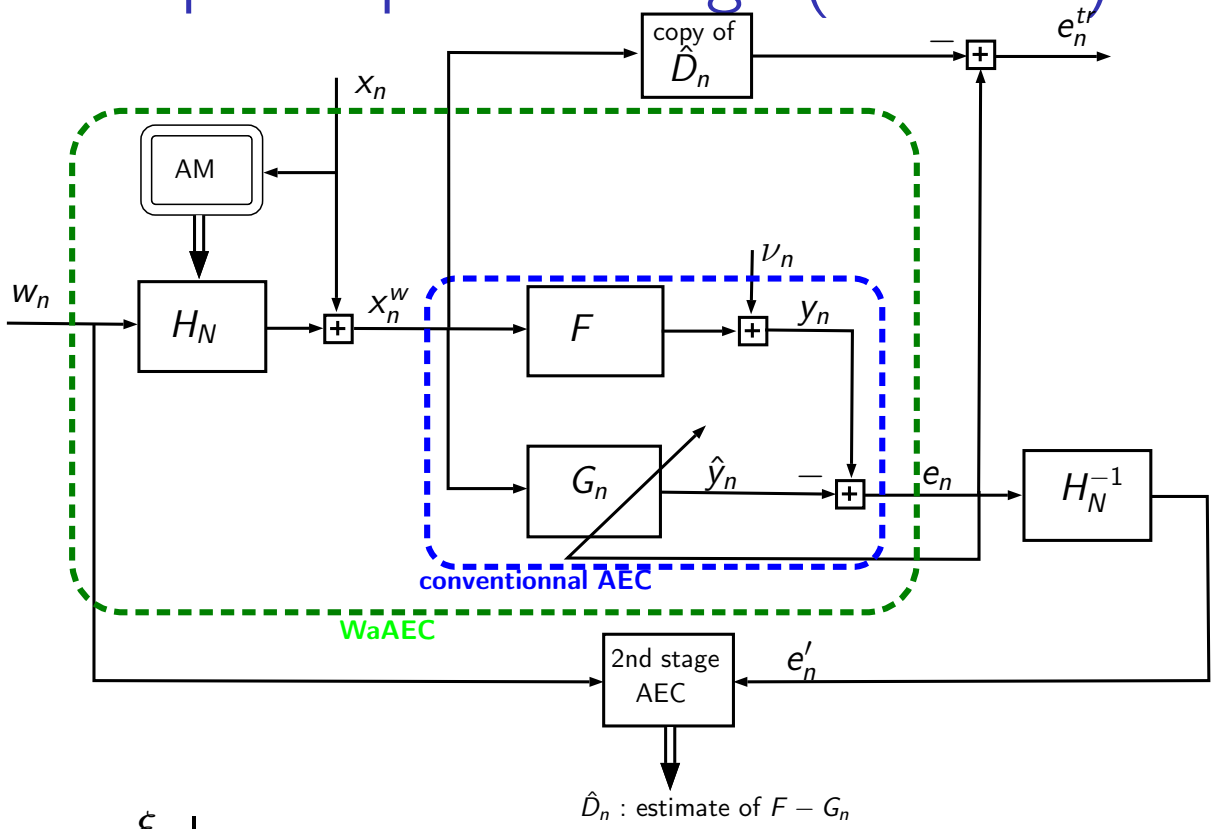


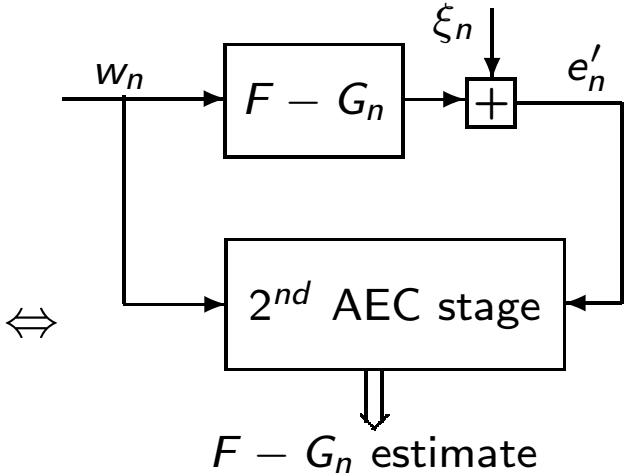
- 1ère idée : dopage par un bruit inaudible
  S. Larbi et M. Jaïdane, Audio watermarking : A way to stationarize audio signals. IEEE Trans. Signal Processing, 2005
  → watermark aided AEC (WaAEC) : accélère convergence et réduit l'écho résiduel de 2 à 5 dB en régime permanent

- 2ème idée : le bruit inséré a toutes les bonnes propriétés...
  Pourquoi ne pas piloter l'identification par le bruit seul ?

- Difficulté : extraire de l'écho du signal bruité l'écho du bruit w seul

Annuleur d'écho piloté par tatouage (WdAEC)





$\Leftrightarrow$

- $\xi_n = \left(\underbrace{(f - g_n)}_{d_n} * x_n + \nu_n \right) * h_N^{-1}$
- Écho résiduel :
  $e_n^{tr} = [d_n - \hat{d}_n] * x_n^w + \nu_n$
  au lieu de $d_n * x_n^w + \nu_n$

Implémentations du 2nd étage

AEC piloté adaptativement par le tatouage (A-WdAEC)

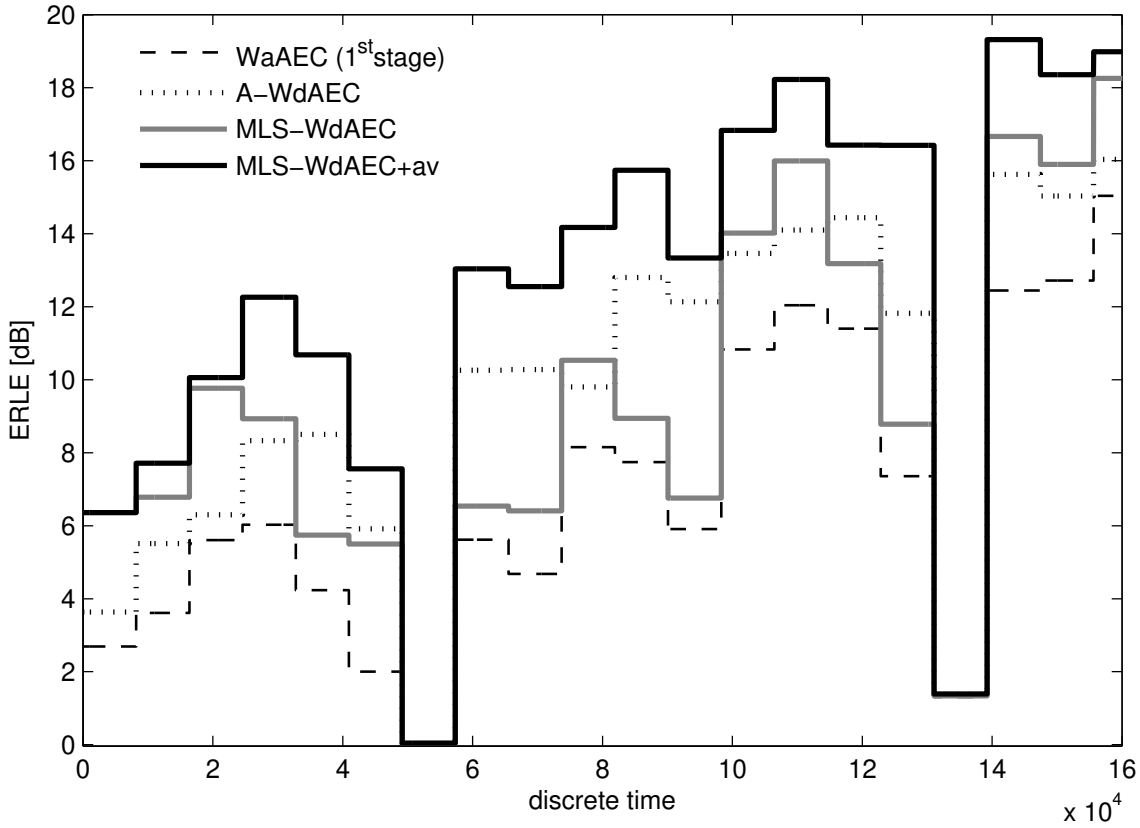
- w_n : bruit blanc gaussien de variance unité.
  2nd étage = filtre adaptatif selon l'algorithme NLMS
- On démontre que
  vitesse de convergence du second étage $\gg$ celle du premier
- **Régime permanent** : étude théorique ne permet pas de conclure
  mais, empiriquement, 2nd étage réduit l'écho de 10 dB

AEC piloté par MLS (MLS-WdAEC)

- w_n : séquences à longueur maximale (MLS)
  2nd étage = corrélateur
- Performances d'identification dépendent de longueur L des MLS
  Amélioration en moyennant sur plusieurs L-périodes l'écho de référence e'_n avant l'intercorrélation
- Simulation : 2nd étage réduit écho de 5 dB, voire 10 dB avec moyennage
  (conditions exp : $L = 8191$, RI de voiture longueur 200, RSB = 30 dB)

Performances comparées en régime permanent

- réponse impulsionnelle de voiture de longueur 200
- RSB de 30 dB
- modélisation exacte pour les deux étages



▶ Écho résiduel réduit de 10 dB avec MLS-WdAEC moyenné

Discussion

Performances du 2nd étage atteintes dans des conditions difficiles

- signal w_n pilotant le 2nd étage mis à zéro pendant la moitié du temps
- bruit d'observation ξ_n puissant et non stationnaire
- réponse impulsionnelle d_n à identifier variable dans le temps

Plan

Mesures objectives de netteté du son

Limites de l'état de l'art :

- confusion netteté-intelligibilité
- traitement séparé de la parole et de la musique
- méthodes non-intrusives fondées sur des techniques d'apprentissage
- netteté vue comme non-altération par une distorsion externe
  i.e. mesure la qualité du canal de transmission.

Manque
une **mesure objective de la netteté intrinsèque** d'un son
sans référence à un original supposé pur
indépendante de son contenu haut-niveau (texte ou musique)

Un détour par le traitement d'images

Une image

La norme de son gradient

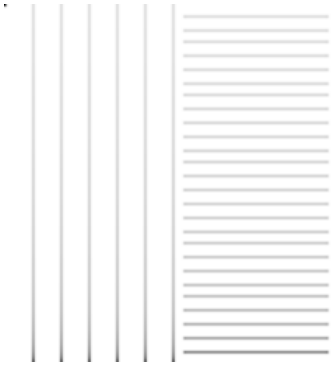
- Image nette $\sim$ gradient parcimonieux $\sim$ variation totale (TV) faible
- Convoluer une image par un bruit blanc augmente plus sa TV si elle est nette que si elle est floue ou bruitée

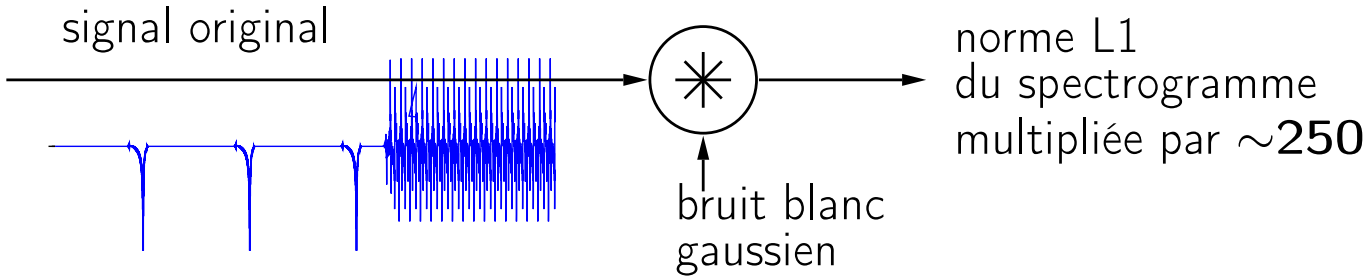
  G. Blanchet and L. Moisan. An explicit sharpness index related to global phase coherence, ICASSP 2012

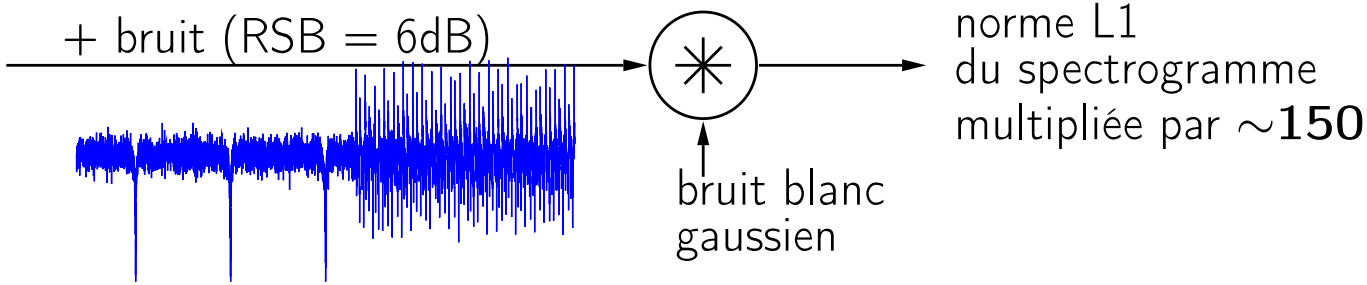
- **Sharpness Index (SI)**
  = sensibilité de la TV à la convolution par un bruit blanc
- **Comment transposer le SI aux signaux audio ?**

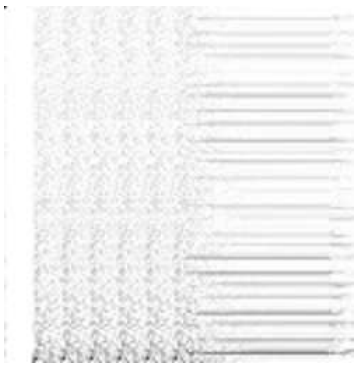
Transposition du Sharpness Index à l'audio

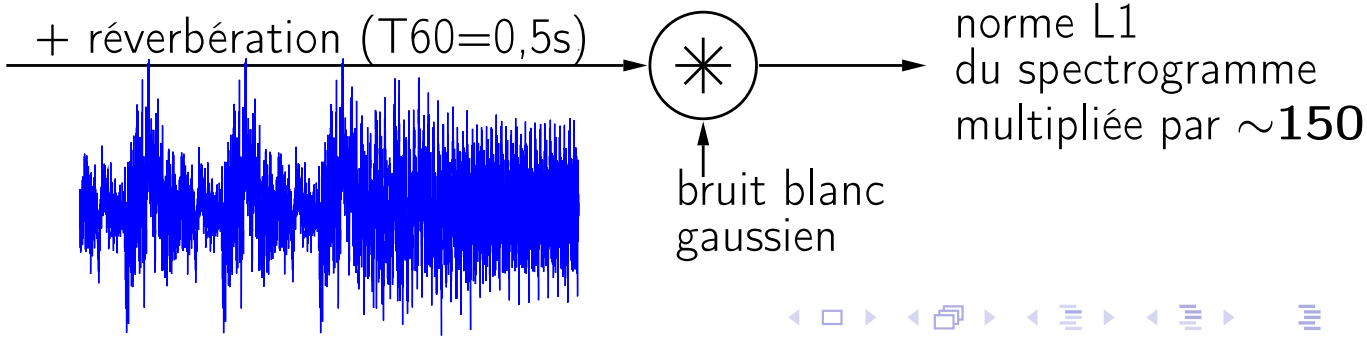
spectrogramme | signal temporel











Non-Intrusive Audio Clarity : définition

NIAC = sensibilité de la parcimonie du spectrogramme $S(f,t)$ à la convolution de s par un bruit blanc gaussien

- Parcimonie de $S(f,t)$ mesurée par $\|S\|_1 = \sum_{f,t} |S(f,t)|$
- Soit $s' = s * w$ avec w bruit blanc $\sim \mathcal{N}(0, \sigma_w^2)$, S' son spectrogramme
- Nous définissons la NIAC comme :

$$\mathcal{C}(s) \triangleq -\log \mathrm{Prob}\left[\|S'\|_1 \leq \|S\|_1\right]$$

- Sous l'hypothèse que $\|S'\|_1$ est gaussienne,

$$\mathcal{C}(s) = -\log\left(\Phi\left(\frac{\mathrm{E}[\|S'\|_1] - \|S\|_1}{\sqrt{\mathrm{Var}[\|S'\|_1]}}\right)\right)$$

où $\Phi(\cdot)$ = queue d'une distribution normale.

Calcul de la NIAC

$\mathrm{E}[\|S'\|_1]$ et $\mathrm{Var}[\|S'\|_1]$ calculées à partir de

$$\Gamma_S(f, f', \tau) \triangleq \mathfrak{T}[\mathcal{R}_{s,\tau}(n, n')]$$

avec

- $\mathfrak{T}$ = version 2D de la transformée utilisée pour le spectrogramme
- $\mathcal{R}_{s,\tau}(n, n') \triangleq \underbrace{R_s(\tau + n - n')}_{\text{auto-corrélation de } s \text{ (finie et déterministe)}} h(n)h(n')$

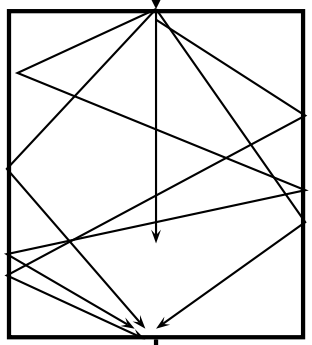
Très peu de paramètres à fixer :

- Spectrogramme : $\mathfrak{T}$, fenêtres d'analyse, recouvrement
- Durée d'analyse $T \simeq 400$ms

Expérience

Signaux de parole

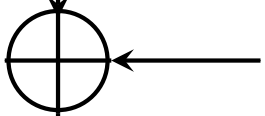
du corpus TIMIT, à 16kHz
16 locuteurs
9 à 18 s de parole chacun



réverbération artificielle

réponse impulsionnelle de salle purement réverbérante

30 temps de réverbération T60 répartis logarithmiquement entre 10ms et 5s



Bruit

bruit blanc gaussien

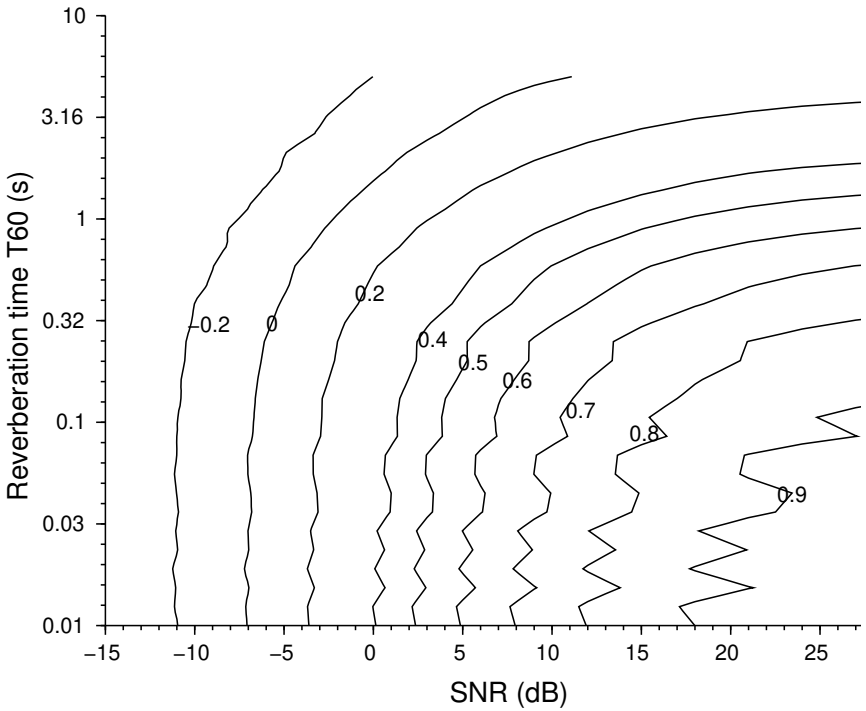
21 valeurs de RSB réparties linéairement entre -30 et +30dB

Globalement, NIAC calculée sur $16 \times 30 \times 21$ signaux

Résultats : comparaison au Speech Transmission Index (STI)

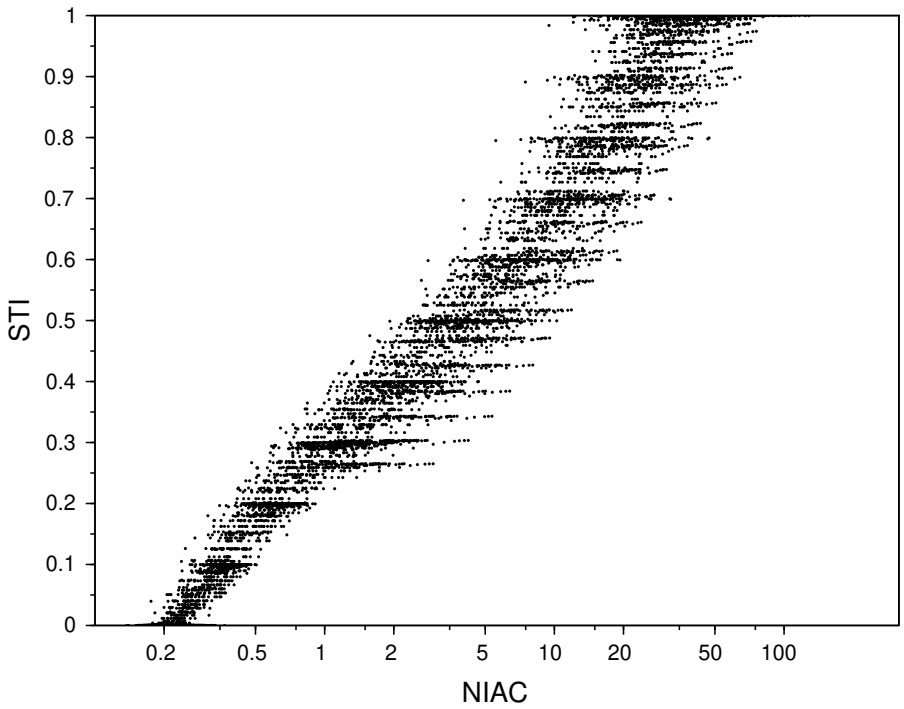
Lignes iso-NIAC dans le plan SNR-T60

Pour chaque condition (SNR,T60), NIAC moyennée sur les 16 locuteurs.



Corrélation logNIAC - STI

- 1 pt = 1 cond. (locuteur,SNR,T60)
- coef corrélation global : 0,96
- 0,94 < coef individuel < 0,98



Utiliser la NIAC pour séparer un mélange audio

Idée : une source séparée est plus nette qu'un mélange

- piloter un algorithme de séparation par la maximisation de la NIAC

Séparation d'un mélange linéaire instantané déterminé

- Mélange $x = As$: s vecteur de p sources, A matrice de mélange $p \times p$
- Posons $y_\alpha = \sum_{i=1}^{p} \alpha_i x_i$, $\alpha = [\alpha_1 \dots \alpha_p]^\top$
- **Extraire une source du mélange = trouver** $\hat{\alpha} \in \arg\max_\alpha \overline{\mathcal{C}}(y_\alpha)$
  où $\overline{\mathcal{C}}$ = moyenne de la NIAC sur plusieurs blocs de durée T.
- Calcul $\mathcal{C}(y_\alpha)$ repose sur :

$$\Gamma_Y(f, f', \tau) = \sum_{1 \le i,j \le p} \alpha_i \alpha_j \Gamma_{x_i x_j}(f, f', \tau),$$

où

  - $\Gamma_{x_i x_j}(f, f', \tau) \triangleq \mathfrak{T}[\tilde{R}_{x_i x_j, \tau}(n, n')]$
  - $\tilde{R}_{x_i x_j, \tau}(n, n') \triangleq R_{x_i x_j}(\tau + n - n') h(n) h(n')$

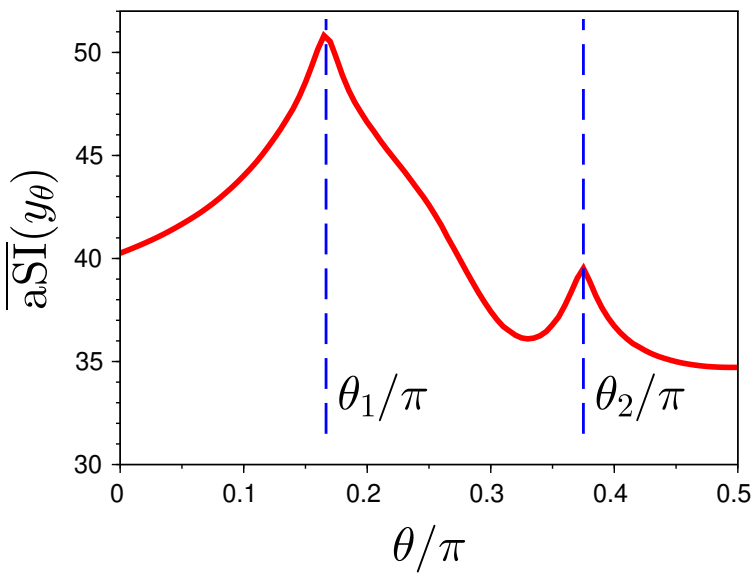
- Les $\Gamma_{x_i x_j}(f, f', \tau)$ sont calculés une seule fois avant l'optimisation

Exemples : séparation de mélanges stéréo de 2 sources

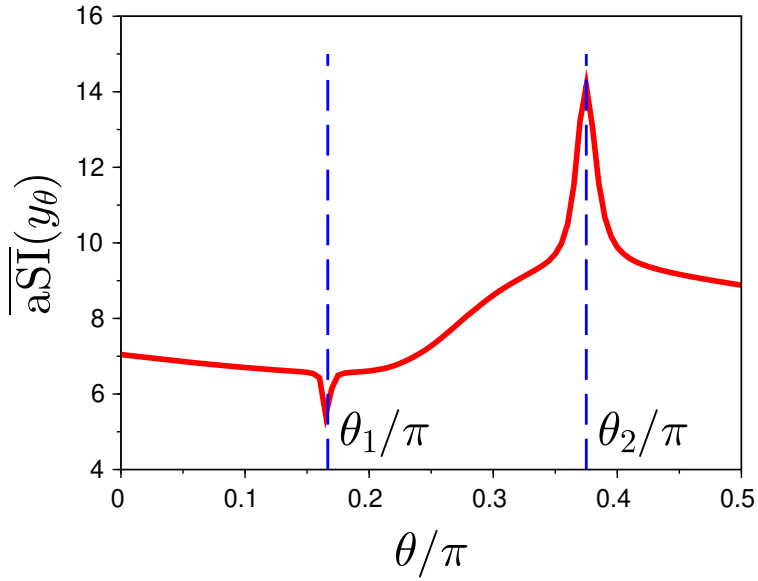
Soit mélange tel que coefficients d'extraction $(\alpha_1, \alpha_2) = (\sin\theta, -\cos\theta)$ et extraction exacte d'une source pour $\theta \in \{\theta_1, \theta_2\}$

NIAC du signal extrait y_θ en fonction de θ :

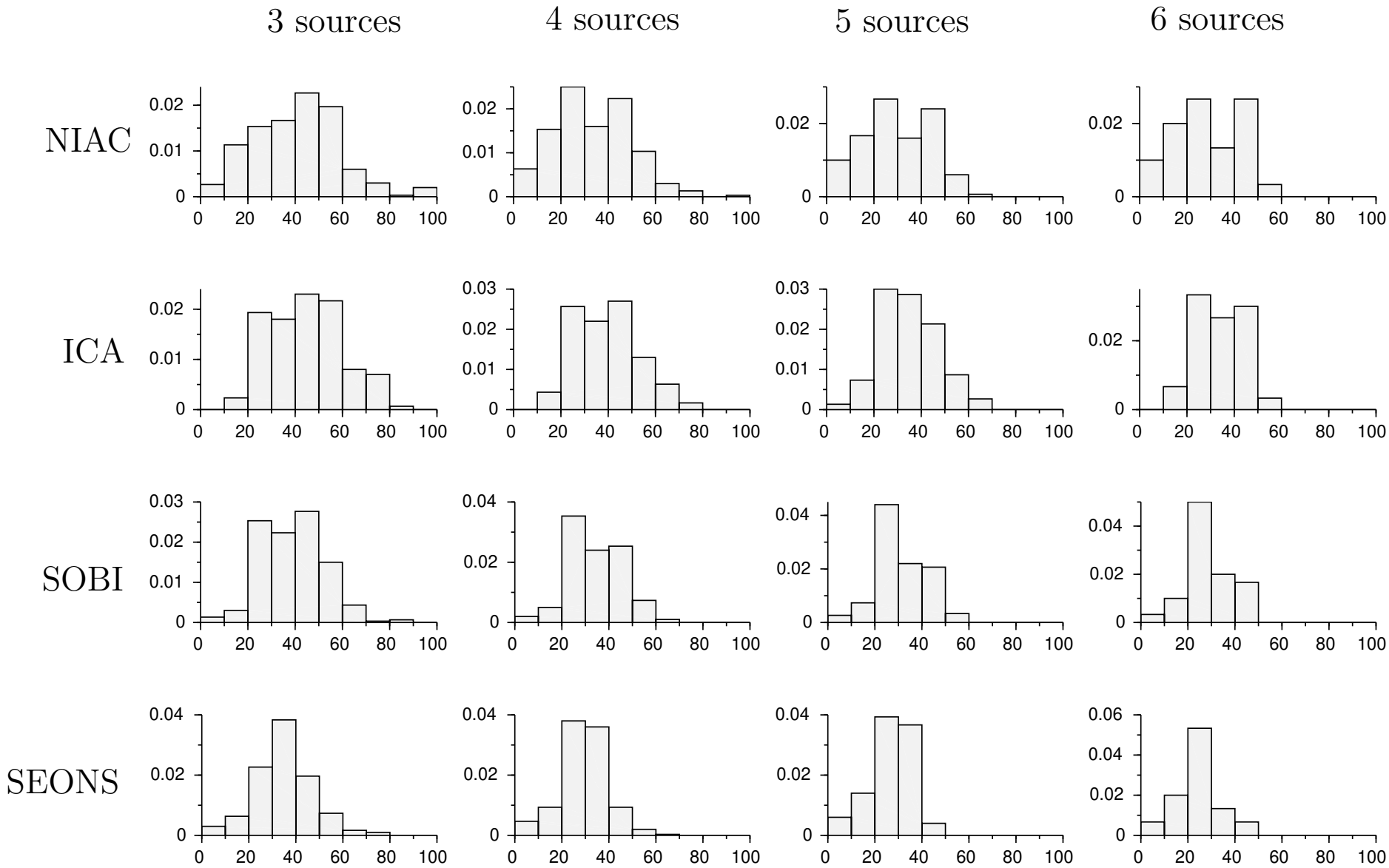
Mélange de 2 voix parlées



Voix chantée + piano



Séparation de mélanges d'un corpus → histogrammes des SIR



En plus...

Pas d'hypothèse d'indépendance ni de non-gaussianité.

Exemple : mélange = voix chantées s et $s' = \text{signe}(s)\sqrt{|s|}$

Démo NIAC :

- ▶ mélange chant-$\sqrt{\text{chant}}$
- Séparation par FastICA : ▶ chant SDR=18dB ; SIR=18dB
- Séparation par Max NIAC : ▶ chant SDR=30dB ; SIR=79dB

Conclusion sur la NIAC

- **Mesure de netteté :**
  - très corrélée au STI sans être intrusive
  - aucun apprentissage ou réglage fin de paramètres
  - non spécifique à la parole ou la musique
  - mesure qualité intrinsèque du signal audio, $\forall$ canal de transmission
- **Séparation de sources** fondée sur la NIAC
  - converge rapidement
  - robuste à l'initialisation de l'algorithme.
  - robuste à la gaussiannité des sources et à leur dépendance
  - supporte une matrice de mélange mal conditionnée

G. Mahé, G. Suzumura, L. Moisan et R. Suyama. A non intrusive audio clarity index (NIAC) and its application to blind source separation. Signal Processing, 2022

Plan

Bilan

Du bruit pour...

informer

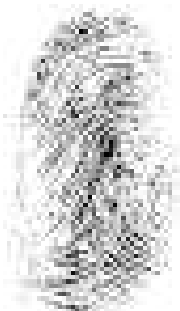
codage audio

séparation de sources

doper

théorème de quantification

codage de canal

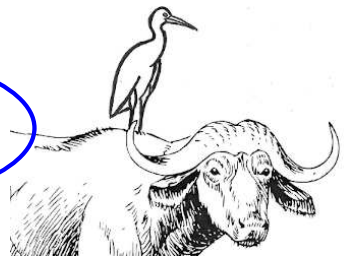


qualité audio

révéler

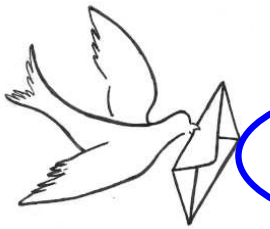
identification de système

témoigner



Ce dont je suis le plus satisfait

Du bruit pour...



informer

codage audio

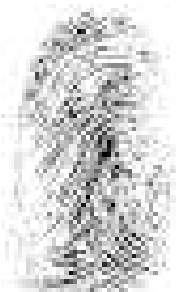
séparation de sources

doper

théorème de quantification

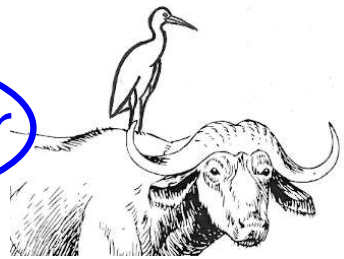
codage de canal

identification de système



qualité audio

révéler



témoigner

Ce dont je suis le plus satisfait

Du bruit pour...

informer

doper

théorème de quantification

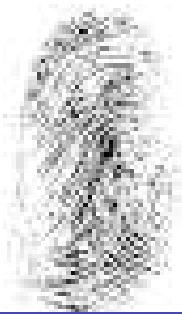
codage audio

codage de canal

séparation de sources

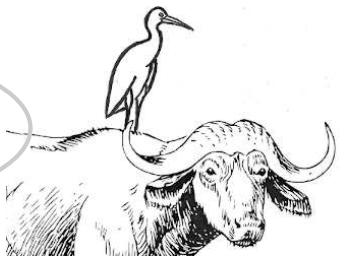
qualité audio

identification de système



révéler

témoigner



Ce dont je suis le plus satisfait

Du bruit pour...

informer

doper

codage audio

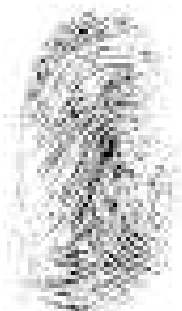
théorème de quantification

codage de canal

séparation de sources

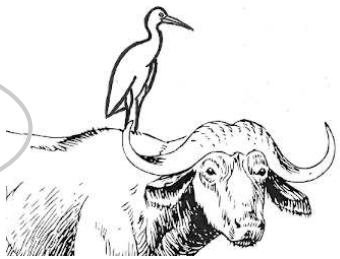
qualité audio

identification de système



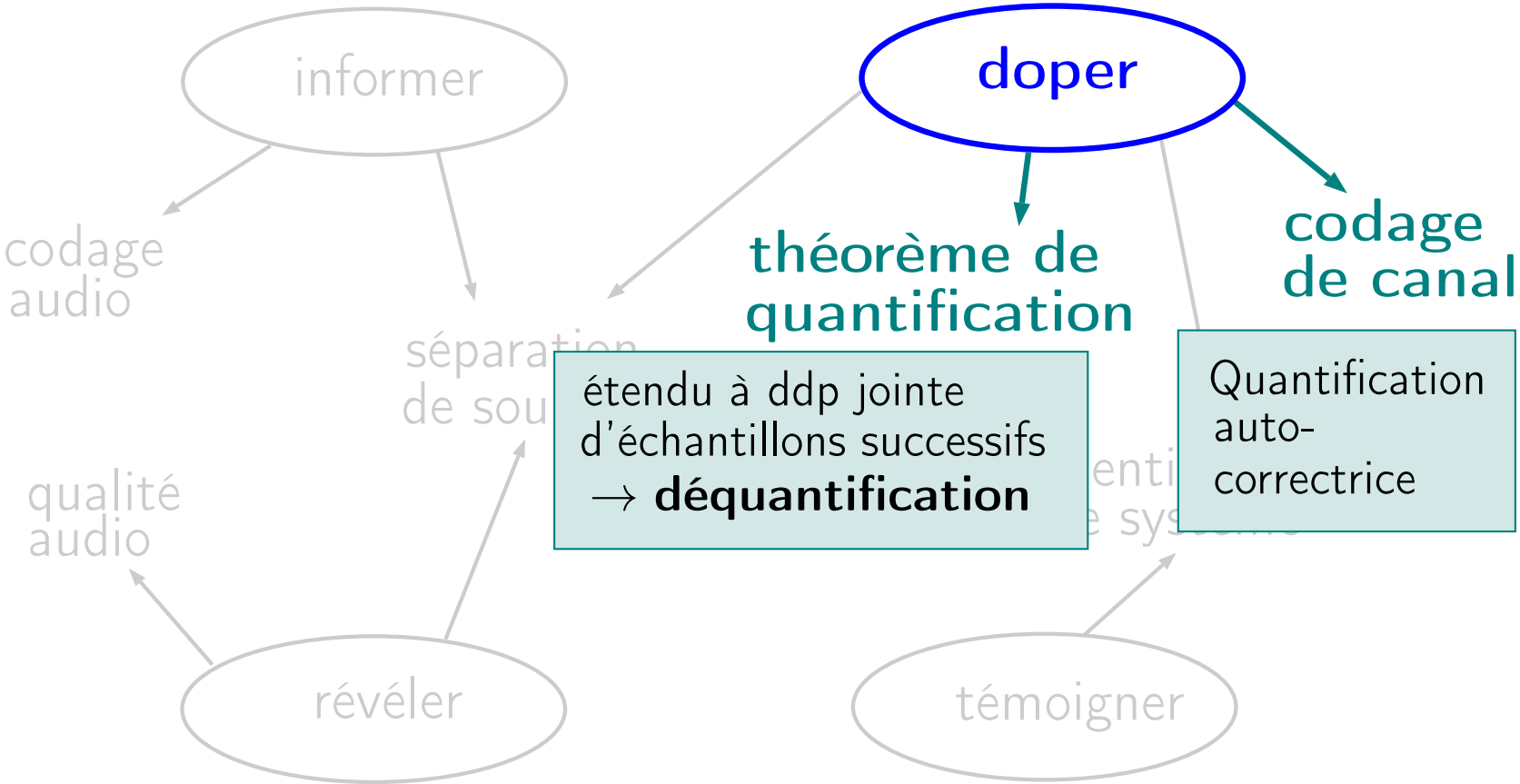
révéler

témoigner



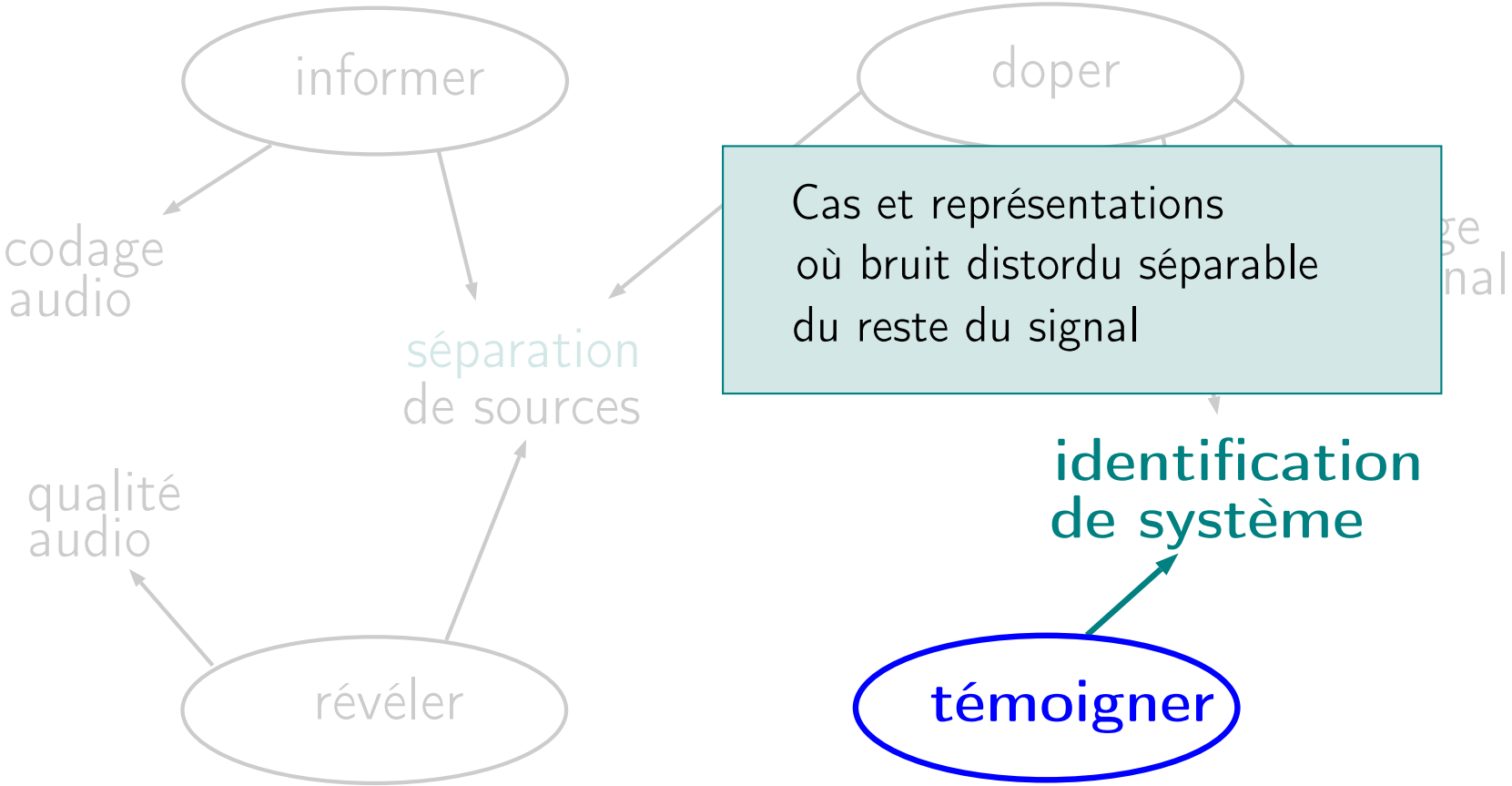
Ce que je souhaite approfondir

Du bruit pour...



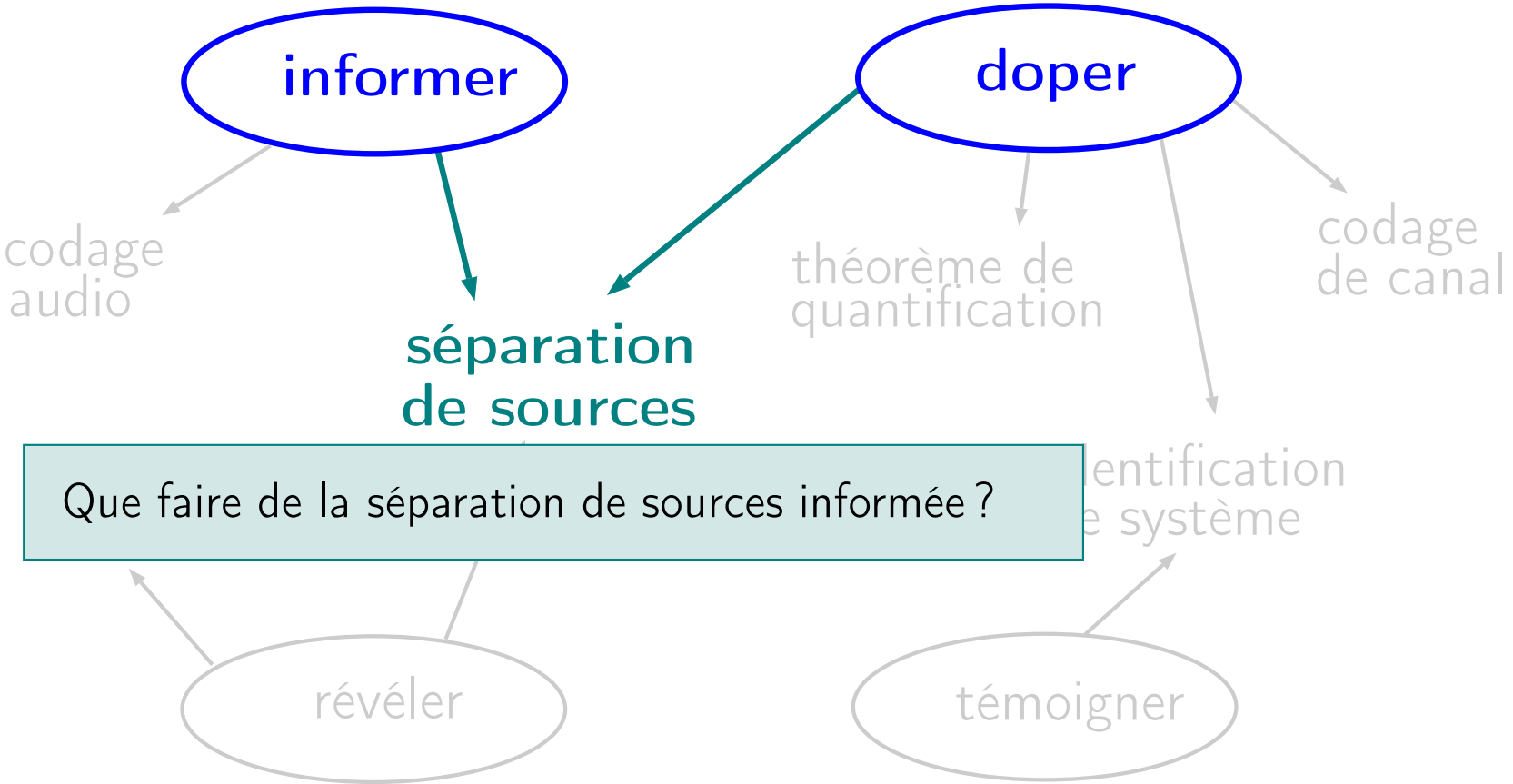
Ce que je souhaite approfondir

Du bruit pour...



Ce que je souhaite approfondir

Du bruit pour...

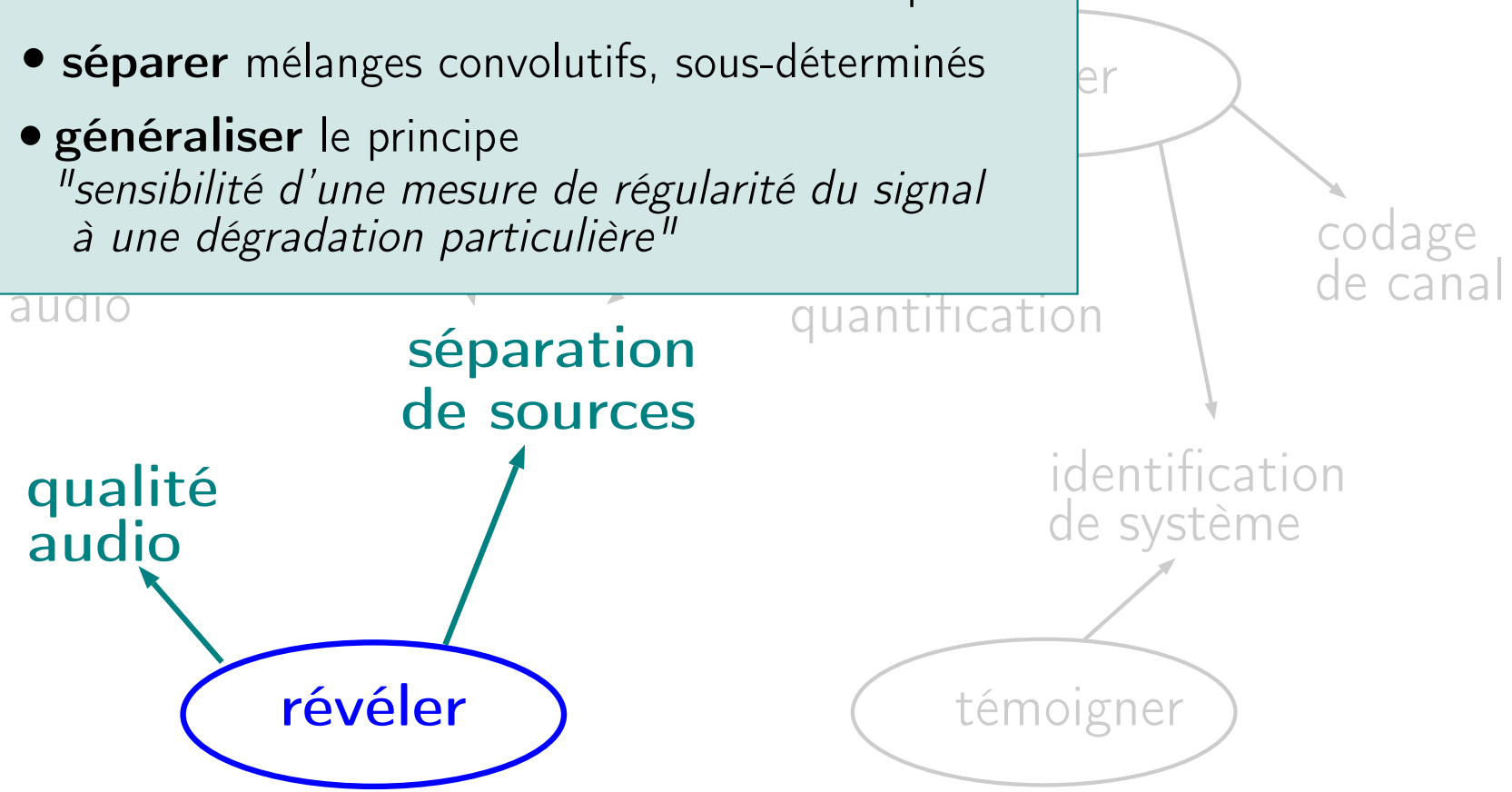


Ce que je souhaite approfondir

Du bruit pour...

NIAC

- **valider** comme mesure de netteté intrinsèque
- **séparer** mélanges convolutifs, sous-déterminés
- **généraliser** le principe
  "sensibilité d'une mesure de régularité du signal à une dégradation particulière"



Des sujets et que je serais curieux d'aborder

- Communication audio au-delà de l'humain
- Canaux de communication exotiques et cachés

Du « *haha !* » au « *ahah...* »

Plan

AEC piloté adaptativement par le tatouage (A-WdAEC)

- w_n : bruit blanc gaussien de variance unité.
- 2nd étage = filtre adaptatif selon l'algorithme NLMS
- **vitesse de convergence** dépend du conditionnement de
  - $\mathrm{R}_{\mathbf{xw}}(n) = \mathrm{E}\left[\frac{X_n^w (X_n^w)^t}{\|X_n^w\|^2}\right]$ pour le premier étage
  - $\mathrm{R}_{\mathbf{w}}(n) = \mathrm{E}\left[\frac{W_n (W_n)^t}{\|W_n\|^2}\right]$ pour le second étage
  - w_n est blanc alors que x_n^w est fortement coloré,
  - $\rightarrow$ vitesse de convergence du second étage $\gg$ celle du premier.
- **Comportement en régime permanent**
  dépend de la puissance instantanée de
  - $\mu\nu_n \frac{X_n^w}{\|X_n^w\|^2}$ pour le premier étage
  - $\mu^w \xi_n \frac{W_n}{\|W_n\|^2}$ pour le second étage,
  - Bruit ξ_n plus puissant et moins stationnaire que ν_n,
  - mais $\frac{W_n}{\|W_n\|^2}$ a des variations temporelles plus douces que $\frac{X_n^w}{\|X_n^w\|^2}$
  - Empiriquement : 2nd étage réduit l'écho d'environ 10 dB (conditions exp. : réponse impulsionnelle de voiture de longueur 200, RSB = 30 dB)

Comparaison avec l'état de l'art (1)

- **Approche classique** :
  adapter les algorithmes aux « mauvaises » propriétés du signal
  Ex Joint-Optimized NLMS (JO-NLMS, Paleologu *et al.*, 2015) :
  pas variable et régularisation
  $\rightarrow$ robuste aux variations locales de la puissance du signal et au bruit
- **Notre approche** :
  adapter le signal aux propriétés requises par un algorithme basique, le NLMS

Comparaison avec l'état de l'art (2)

Comparaison efficacité :

- Gain d'ERLE du MLS-WdAEC moyenné par rapport au JO-NLMS : $\delta_{\text{ERLE}} = \text{ERLE}_{\text{MLS-WdAEC}} - \text{ERLE}_{\text{JO-NLMS}}$ dans différentes situations
- $p =$ longueur de la réponse impulsionnelle du canal
- $M =$ longueur de la réponse impulsionnelle du filtre identificateur.