# Le beau noiseur

Gaël Mahé

MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université Paris Cité

LIPADE
Laboratoire d'Informatique PAris DEscartes

Université Paris Cité / Faculté des Sciences
UFR de mathématiques et informatique / LIPADE
45 rue des Saints-Pères - Paris 6e

# 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**

par

**Gaël Mahé**

soutenue publiquement le 9 janvier 2023 devant le jury composé de

| | | |
|---|---|---|
| Laurent Daudet | professeur à l'Université Paris Cité | président |
| Sylvain Marchand | professeur à l'Université de La Rochelle | rapporteur |
| Phillip A. Regalia | directeur de programme à la National Science Foundation | rapporteur |
| Gaël Richard | professeur à Télécom Paris | rapporteur |
| Rosângela Coelho | professeure à l'Instituto Militar de Engenharia de Rio de Janeiro | examinatrice |
| Mériem Jaïdane | professeure à l'École Nationale d'Ingénieurs de Tunis | examinatrice |
| Régine Le Bouquin Jeannès | professeure à l'Université de Rennes | examinatrice |
| Olivier Warusfel | chargé de recherche à l'IRCAM | examinateur |

# Table des matières

# Résumé

Nous montrons comment des distorsions inaudibles du son peuvent faciliter des traitements d'analyse ou de correction de celui-ci. En partant de la notion de tatouage audio, un bruit inaudible prendra ainsi quatre fonctions originales — *informer, doper, témoigner* et *révéler* — au service du traitement du son.

La première des distorsions étudiées est celle du tatouage audio proprement dit, pour lequel plusieurs contributions sont présentées, soit améliorant des techniques existantes par le renforcement de la robustesse ou de l'inaudibilité, soit proposant une nouvelle technique. Le tatouage est utilisé ici pour *insérer dans le son une information utile à un système de traitement en réception d'une chaîne de transmission* : soit une information sur les propriétés du signal, soit l'information manquante pour traiter le signal connaissant la version dégradée de celui-ci. Nous montrons que ce *tatouage réflexif* permet la restauration d'un signal audio en sortie d'un canal de communication dégradé dans deux contextes applicatifs : la compression audio et la séparation de sources dans un mélange stéréo.

La notion de tatouage est ensuite étendue pour assigner à cette distorsion du son, non plus le rôle de transmettre une information explicite, mais celui de *modifier les propriétés du signal hôte de manière à faciliter un traitement ultérieur.* Nous montrons que ce *dopage* du signal permet d'améliorer les performances d'algorithmes génériques en facilitant la vérification d'hypothèses statistiques trop fortes sur lesquels reposent idéalement ces derniers. Différentes applications génériques en traitement du son sont revisitées : l'identification de systèmes, la quantification, la séparation de sources et le débruitage.

Lorsque le traitement d'un signal issu d'un canal de communication dégradé repose sur l'identification de ce canal, un bruit ajouté au signal peut faciliter cette identification dès lors que l'on peut comparer sa version originelle avec celle ayant subi les mêmes altérations que son signal hôte. Cette idée d'un *bruit témoin des altérations du signal* est validée dans le cas de l'identification de systèmes linéaires, notamment pour l'annulation d'écho acoustique.

Enfin, nous proposons de bruiter le signal virtuellement pour l'analyser et le traiter : le bruit ajouté permet de mesurer la netteté d'un son et d'utiliser cette mesure comme un critère efficace de séparation de sources. Après avoir informé, dopé, témoigné, *le bruit agit comme un révélateur du signal.*

# Abstract

We show how inaudibly distorting an audio signal can make further analysis or correction processing easier. Starting from the notion of audio watermarking, an inaudible noise will then play four original functions — *informing, doping, witnessing*, and *revealing* — dedicated to sound processing.

The first studied distortion is that of audio watermarking itself, for which several contributions are presented, that either enhance existing techniques by reinforcing the watermark robustness or inaudibility, or propose a new technique. Here, watermarking is used to *insert information into the sound that is useful to a processing system at the receiving part of a transmission chain* : either information on the signal properties, or the missing information to process the signal knowing the impaired version of it. We will show that this *reflexive watermarking* enables to restore an audio signal at the output of a corrupted communication channel in two application contexts : audio compression and source separation from a stereo mixture.

The notion of watermarking is then extended to assign to this distortion of the sound, no longer the role of transmitting explicit information, but that of *modifying the properties of the host signal so as to make further processing easier.* We show that this signal *doping* enables to enhance the performance of generic algorithms by helping to verify too strong statistical assumptions they ideally rely on. Various generic audio processing applications are revisited : system identification, quantization, source separation and denoising.

When processing a signal from a degraded communication channel relies on the identification of the this channel, a noise added to the signal can help this identification if one can compare its original version with the one having undergone the same alterations as its host signal. This idea of a *noise witness of signal alterations* is validated in the case of linear system identification, especially for acoustic echo cancellation.

Finally, we propose to add a virtual noise to the signal for analyzing and processing purposes : the added noise enables measuring the sound clarity and using this measure as an efficient source separation criterion. After having informed, doped, and been witness, *the noise reveals the signal.*

# Introduction

Ma thèse au Centre National d'Étude des Télécommunications (CNET [1]), sur la correction des distorsions spectrales de la parole dans les réseaux téléphoniques, visait à rapprocher le timbre de la voix téléphonique de l'original, tout en contrôlant les effets du traitement sur le bruit de quantification. Ce sujet appliqué ne s'inscrivait pas dans un sous-domaine précis du traitement du signal, mais requérait des outils divers issus du traitement du signal (codage de source, quantification...) et d'autres domaines (classification, psychoacoustique... ). Si j'ai abordé des sujets différents après la thèse, une continuité apparaît non seulement *via* certains outils théoriques, mais aussi dans la démarche consistant à « piocher » dans l'ensemble du traitement du signal pour répondre à une question non spécifique à un sous-domaine de celui-ci, comme on le verra par la suite.

À mon arrivée au CRIP5 (devenu LIPADE en 2009) fin 2002, au sein de l'équipe InfoCom [2], j'ai orienté ma recherche vers un des thèmes récemment développés par cette équipe, le tatouage (ou watermarking) audio. Celui-ci consiste à insérer une information dans un signal (son, image, vidéo...) par une altération imperceptible de ce signal. Ce choix était influencé par le caractère ludique du tatouage, par l'impression d'un champ de recherche largement ouvert et inexploré et par mon parcours antérieur : de formation télécom avec une forte dominante en communications numériques, je m'étais tourné vers l'audio au cours de ma thèse ; le tatouage audio avait l'avantage de mêler ces deux « cultures », avec un état de l'art dont les manques étaient nets : les questions de sécurité et de théorie de l'information étaient privilégiées, au détriment des aspects proprement audio.

Les techniques de tatouage sont toutefois une niche, dont je suis rapidement sorti pour réfléchir plutôt à l'utilisation du tatouage. Alors que les recherches en tatouage étaient souvent tournées vers des objectifs sécuritaires (*Digital Rights Management*), j'ai proposé une approche guidée par une nouvelle utilisation du tatouage. Il s'agit de définir des tatouages intimement liés au signal hôte, facilitant la restauration, l'analyse ou la manipulation de ce signal à la sortie d'un canal de communication dégradé. Cette approche a fait l'objet du projet WaRRIS [3] à partir de 2006, mené en collaboration avec l'Unité Signaux et Systèmes (U2S) de l'ENIT [4]. Elle se décline en trois concepts :

1. devenu France Télécom R&D, puis Orange Labs
2. dont les travaux portaient sur l'identification de systèmes, les algorithmes adaptatifs, les méthodes algébriques et le codage audio
3. Watermarking Réflexif pour le Renforcement des Images et des Sons : projet ANR jeunes chercheurs que j'ai porté, financé de 2006 à 2010.
4. École Nationale d'Ingénieurs de Tunis. L'U2S est devenu récemment le Laboratoire Signals and Smart Systems (L3S).

— le *tatouage d'accentuation* ou *tatouage réflexif* est conçu comme une mémoire porteuse d'informations sur le signal originel, ce qui permet de restaurer la qualité audio d'un son issu d'un canal de communication dégradé ;

— le *tatouage dopant* consiste à modifier de manière imperceptible les propriétés d'un signal pour faciliter des traitements ultérieurs d'analyse ou de correction ;

— le *tatouage témoin* s'imprègne des altérations subies par le signal hôte, de sorte que la comparaison entre le tatouage altéré et le tatouage initial connu permet d'estimer ces altérations et, partant, de les inverser.

Ces différentes approches existaient dans l'état de l'art, mais de manière marginale et dispersée. Ce projet visait à les systématiser et à en démontrer la pertinence générale via des études de cas classiques touchant tout le traitement du signal audio : identification de systèmes, séparation de sources, codage audio, quantification... Il a guidé mes travaux jusqu'à aujourd'hui.

D'autres thèmes et d'autres collaborations sont apparus au fil des résultats, des rencontres, des appels à projets et de l'évolution de mon environnement de recherche. Une collaboration avec les universités de Campinas et d'ABC, au Brésil, m'a permis de travailler sur la séparation de sources, dans le cadre de projets CAPES [5]-COFECUB [6] (2013-2017 et 2020-2023). Le projet européen ICityForAll (2011-2015) a été l'occasion de m'intéresser à la détection des alarmes, à la saillance auditive et à la mesure de la netteté du son. Tout en visant une application industrielle, nous avons renforcé l'outillage théorique de nos travaux par une collaboration avec le laboratoire Mathématiques Appliquées à Paris 5 (MAP5) sur la mesure de netteté.

Sans que ce fût toujours prémédité, tous ces nouveaux thèmes s'inscrivent dans l'esprit du projet WaRRIS ou le prolongent [7]. L'idée générale qui préside aux travaux présentés dans ce mémoire est la suivante : **concevoir des distorsions inaudibles du son qui facilitent des traitements d'analyse ou de correction**.

Mes contributions aux **techniques de tatouage audio** sont présentées dans le chapitre 1 : il s'agit d'améliorations de techniques existantes (robustesse ou inaudibilité du tatouage) et d'une proposition d'une nouvelle technique.

Comment le tatouage pourrait-il faciliter un traitement d'analyse ou de correction du son ? La première idée est d'insérer dans le signal une information utile à un système de traitement en réception d'une chaîne de transmission : soit une information sur les propriétés du signal original, soit l'information qui manque pour traiter le signal connaissant la version dégradée de celui-ci. C'est l'objet du chapitre 2, avec des applications au codage audio et à la séparation de sources : **du bruit pour informer - le tatouage réflexif**.

Puis nous nous sommes rendu compte qu'un tatouage peut être utile au traitement du son non seulement par l'information binaire qu'il transporte, mais aussi en ce qu'il transforme les propriétés du signal hôte. Le chapitre 3 développe cette idée d'un tatouage dopant, au service d'applications telles que l'identification de sys-

5. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

6. Comité Français d'Évaluation de la Coopération Universitaire et Scientifique avec le Brésil

7. Nous n'évoquerons pas dans ce mémoire une autre collaboration importante, mais en dehors du fil conducteur des autres travaux : celle avec le laboratoire Cognac-G de l'Université Paris Cité (Cognition and Action Group, aujourd'hui Centre Borelli) sur l'intelligence collective sensorimotrice puis sur la détection du sourire, occasion d'appliquer les outils du traitement du signal dans un contexte plus inter-disciplinaire. Voir documents annexes.

tèmes, la quantification, la séparation de sources ou le débruitage : **du bruit pour doper le signal - le tatouage dopant**.

Dans les applications classiques du tatouage, il en est une où celui-ci est utile par son absence : en disparaissant lors d'une manipulation illicite du signal, il permet de détecter celle-ci. Le chapitre 4 montre une extension de cette idée avec le tatouage témoin, qui subit les mêmes distorsions que son signal hôte, tout en possédant des propriétés qui facilitent l'identification de ces distorsions par comparaison entre le tatouage originel et sa version distordue : **le bruit témoin des altérations du signal - le tatouage témoin**.

Dans les tatouages dopant et témoin, le tatouage n'est plus nécessairement un tatouage au sens usuel de message binaire inséré via un bruit inaudible ajouté au signal ou une modification imperceptible du signal : il peut se limiter à un bruit spécifiquement conçu pour l'application visée. Dans le chapitre 5, nous poursuivons cette éloignement du tatouage : le bruit n'est plus inséré de manière inaudible, mais ajouté virtuellement, par le calcul, pour mesurer la netteté du son. Le bruit agit alors comme un **révélateur du signal**.

*Les acronymes d'institutions auxquels le texte fait référence sont explicités p.121. Pour chaque sujet, nous avons indiqué le ou les projet(s) dans le cadre duquel ou desquels il a été traité. Les projets référencés sont précisés p.123.*

# Chapitre 1

# Tatouage : le jeu de l'acrostiche

**Sommaire**



Le tatouage (ou watermarking) consiste à insérer une information dans un document ou un signal (texte, son, image, vidéo...) par une altération imperceptible de cet objet. L'information est alors disponible pour qui connaît son existence et dispose de la clé pour l'extraire. Bien que ce procédé soit vieux comme le monde (que l'on songe aux acrostiches), c'est dans les années 1990 qu'il s'est développé en tant que sujet de recherche [17], favorisé par une demande industrielle : la diffusion numérique des productions audio(visuelles) facilitait le piratage, ce qui a suscité la demande de nouveaux moyens de protection du copyright. Le tatouage a ainsi de nombreuses applications sécuritaires : insertion de marques de copyright ; insertion d'un identifiant unique d'exemplaire permettant le traçage des copies pirates ; tatouage fragile anti-falsification, se détruisant à la moindre manipulation du document. Une autre classe d'applications, plus proche de la stéganographie historique, consiste à utiliser l'objet tatoué comme un canal parallèle de transmission, ce qui permet d'enrichir son contenu : citons par exemple le projet RNRT Artus [1], dans lequel le tatouage d'un signal de parole par sa traduction en langue des signes permettait l'animation d'un avatar pour les malentendants.

Le tatouage doit respecter plusieurs contraintes contradictoires : il doit être imperceptible, résister aux manipulations de l'objet support (à l'exception du tatouage fragile évoqué précédemment) et assurer un débit d'information suffisant pour l'application visée. Le compromis entre ces trois contraintes dépend de la classe d'application visée : tandis que les applications sécuritaires privilégient la robustesse (y compris à des attaques visant explicitement la destruction du tatouage), les applications d'enrichissement de contenu nécessitent surtout des débits d'insertion élevés.

L'émergence du tatouage audio comme sujet de recherche s'est accompagné d'une grande variété de propositions de techniques d'insertion de l'information. On peut classer celles ci en deux familles : dans le *tatouage additif*, l'information est codée sous forme d'un signal particulier que l'on ajoute au signal audio et qui doit être masqué par celui-ci (voir 1.4) ; dans le *tatouage substitutif* (1.1, 1.2, 1.3), l'information est insérée en modulant de manière inaudible certains paramètres du signal.

Ces techniques ont en commun d'exploiter l'imperfection de l'oreille humaine [21] pour que le tatouage soit inaudible. Le tatouage additif s'appuie sur les propriétés de masquage fréquentiel (tatouage par étalement de spectre [11]) ou temporel (tatouage par insertion d'écho [7]), tandis que le tatouage substitutif joue sur les variations *justes audibles* de fréquence, de phase, d'amplitude... Notons que la littérature du domaine, dans les années 1990 et 2000, étudiait le tatouage soit de manière très empirique, soit avec des fondements théoriques liés à la théorie de l'information, sans référence à la psychoacoustique, la mesure de la distorsion créée par le tatouage se limitant souvent à l'erreur quadratique moyenne. Nous avons abordé différentes techniques en nous efforçant de fonder celles-ci sur des bases psychoacoustiques.

Le choix des techniques étudiées s'est fait au hasard des rencontres. Ce sont les résultats d'un stage de DEA fournis par mon directeur de thèse, Jean-Marc Boucher, qui ont suscité ma curiosité pour le tatouage par filtrage passe-tout (section 1.1), dont le principe m'apparaissait d'une puissante simplicité. L'étude de la quantification par modulation d'index (section 1.2) est née de discussions avec Cléo Baras à l'ENST et du fait que cette technique soulevait des questions sur la quantification similaires à celles abordées dans ma thèse. L'idée du tatouage par modulation des trajectoires TVAR (section 1.3) est venue après une étude de l'identification de modèle TVAR par filtrage particulaire avec Monia Turki (ENIT), qui s'est avérée fastidieuse et à l'issue de laquelle notre contribution possible me semblait limitée si nous ne sortions pas du cadre. Enfin, j'ai pu bénéficier du travail réalisé précédemment entre le LI-PADE, l'U2S (ENIT) et le LTCI (ENST), *via* les thèses de Leandro Gomes [5], Sonia Larbi [12] et Cléo Baras [2], sur le tatouage par étalement de spectre (section 1.4).

Nous avons étudié ces techniques de tatouage d'abord pour elles-mêmes, puis dans la perspective d'utiliser le tatouage à des fins d'analyse ou de correction du signal hôte, ce qui poussait à rechercher le meilleur débit d'insertion et à considérer le tatouage en lien avec le traitement du signal visé.

## 1.1 Tatouage par filtrage passe-tout

**Collaboration :** Sonia Larbi, U2S, ENIT

**Encadrement :** Salsabil Besbes, projet de fin d'études d'ingénieur ENIT (2007), co-encadré avec Sonia Larbi

**Projets et financements :** WaRRIS et AUF (2007-2008)

Le tatouage par filtrage passe-tout, initialement proposé par [20], repose sur une idée simple : si l'on filtre un signal par un filtre passe-tout de fonction de transfert

$$H(z) = \frac{-a + z^{-1}}{1 - az^{-1}}, \tag{1.1}$$

alors la transformée en $z$ du signal résultant est nulle en $1/a$ et infinie en $a$. Ainsi, en définissant chaque symbole de tatouage comme un pôle de filtre passe-tout, la détection se fait aisément en calculant la transformée en $z$ du signal tatoué en chaque symbole ou en chaque inverse de symbole. Ce principe posé, la réalisation se heurte à deux principaux obstacles liés au découpage du signal en blocs successifs d'insertion.

D'une part, cette troncature se traduit par l'annulation du pôle [15], ce qui oblige à ne fonder la détection du tatouage que sur les zéros des filtres.

D'autre part, le déphasage brutal d'un bloc au au suivant se traduit par des discontinuités audibles (clics), tandis que l'adoucissement des transitions entre blocs perturbe fortement la détection du symbole. Les discontinuités entre blocs peuvent être adoucies en initialisant le filtre associé à un bloc par l'entrée et la sortie du filtrage précédent. Mais il faut alors corriger l'effet de ces nouvelles conditions initiales sur la détection, en retranchant au signal tatoué, lors de la détection, la *zero input response* (ZIR) [4]. Cette solution suppose connu le dernier échantillon du bloc précédent non-tatoué : pour cela, seul un bloc sur deux est tatoué dans [4].

Nous avons étudié ces méthodes lors du stage de fin d'études d'ingénieur ENIT de Salsabil Besbes en 2007. Des clics audibles subsistent avec la méthode proposée par [4], contrairement à ce qui était annoncé. Pour y remédier, nous avons proposé de filtrer chaque bloc par un filtre dont le pôle varie progressivement de 0 à sa valeur maximale sur le premier quart du bloc, puis de cette valeur à 0 sur le dernier quart. Le tatouage est alors détecté sur la moitié centrale du bloc. Nous avons montré qu'il est possible de calculer exactement la ZIR à partir du dernier échantillon du bloc. Pourtant, notre détecteur échoue à détecter correctement le tatouage, ce que nous ne sommes pas parvenus à expliquer.

De nouveaux résultats sur le tatouage par filtrage passe-tout [15] ont été publiés à la même époque. Cependant, pour des débits de tatouage modestes eu égard à nos besoins (jusqu'à 243 bit/s), les taux d'erreur sont assez élevés et les aspects perceptifs semblent partiellement ignorés.

Nos expériences et la littérature laissant peu d'espoir sur la possibilité d'un tatouage haut-débit par filtrage passe-tout avec de solides bases psychoacoustiques, nous avons laissé cette piste de côté.

## 1.2 Quantification par modulation d'index (QIM)

**Collaboration :** Cléo Baras, LTCI, ENST

**Encadrement :** Malek Boujemaa, projet de fin d'études d'ingénieur ENIT (2007)

**Projets et financements :** WaRRIS et AUF (2003-2009)

La quantification par modulation d'index (Quantization Index Modulation, QIM [3]), consiste à définir dans l'espace de représentation du signal autant de sous-ensembles de quantificateurs que de symboles de tatouage et à approcher chaque vecteur du signal par le quantificateur le plus proche associé au symbole à insérer, comme illustré par la figure 1.1. La détection du symbole consiste à identifier à quel sous-ensemble de quantificateurs appartient celui le plus proche du vecteur de signal reçu. La taille

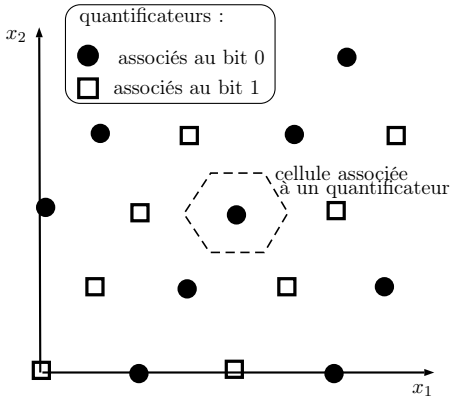



FIGURE 1.1 – Exemple de QIM : insertion d'1 bit dans un vecteur de 2 échantillons $(x_1, x_2)$, par sélection du quantificateur le plus proche appartenant à l'un des deux ensembles de quantificateurs.

et la forme des cellules de quantification détermine le compromis entre l'audibilité du tatouage et sa robustesse aux distorsions.

Une variante de la QIM est la dither modulation (DM), fondée sur le principe de quantification par dithering [10]. Un seul ensemble de quantificateur est utilisé, le tatouage repose sur un vecteur de *dithering* : à chaque symbole $m$ est associé un vecteur $d(m)$. Un vecteur de signal $x$ est tatoué en

$$\tilde{x}_m = \mathrm{q}(x + d(m)) - d(m) \tag{1.2}$$

où q désigne la quantification. En réception, le symbole est détecté dans le vecteur $y$ reçu en recherchant $m$ tel que $y + d(m)$ soit le plus proche de son quantificateur. Le dithering permet d'améliorer le compromis distorsion / robustesse du tatouage.

Ce compromis est encore amélioré par la *spread-transform dither modulation* (STDM), en appliquant la dither modulation à une projection du vecteur signal sur un vecteur ou un sous-espace vectoriel. Le choix de ce(s) vecteur(s) de projection permet, pour un pas de quantification donné (donc à robustesse constante), d'étaler la distorsion sur les composantes du vecteur signal de manière à réduire la distorsion globale. Chen et Wornell ont montré l'optimalité de la STDM du point de vue de la théorie de l'information [3], ainsi que la supériorité de la STDM sur les méthodes additives à étalement de spectre, en terme de compromis débit/robustesse/distorsion.

Nous avons expérimenté la STDM en 2004, en lien avec Cléo Baras [2], puis dans le cadre du projet de fin d'études ENIT de Malek Boujemaa en 2007. Ces expériences montrent que **la STDM permet d'atteindre un débit de tatouage très élevé, proche du kbit/s, tout en en respectant la contrainte d'inaudibilité, mais que le tatouage s'avère peu robuste aux perturbations du canal** : en pratique, les performances sont similaires à celles d'un tatouage additif à étalement de spectre dans le cas d'un canal bruité ou d'une compression MPEG modérée, et nettement inférieures face à d'autres dégradations comme le filtrage passe-bas ou l'insertion d'écho.

Dans [3], Chen et Wornell restent élusifs sur le choix optimal du sous-espace de projection en fonction de critères psychoacoustiques, considérant des critères de type rapport signal à bruit et une approche orientée théorie de l'information. Dans notre mise en œuvre, nous avons considéré des vecteurs de projection aléatoires, uniformément égaux à 1 ou alternant 1 et -1, ce qui conduit à un bruit de quantification blanc,

donc sous-optimal pour le compromis débit/robustesse/distorsion. Au cours du projet de fin d'études de Malek Boujemaa, **nous avons donc cherché à reformer spectralement le bruit de quantification.**

Une première idée est de filtrer le signal, avant tatouage, par un filtre de réponse fréquentielle l'inverse du seuil de masquage, puis par l'inverse de ce filtre après tatouage. Le bruit de quantification résultant du tatouage a ainsi une densité spectrale de puissance égale au seuil de masquage moins une certaine marge, réglée par le pas de quantification. Ce procédé nécessite, en réception, de connaître le seuil de masquage du signal avant tatouage. Dans le cas d'un canal sans distorsion, ce seuil peut être estimé à partir du signal tatoué. Mais cette estimation est très sensible au bruit du canal, de sorte que même un rapport signal à bruit de 50 dB augmente catastrophiquement le taux d'erreur de détection.

Nous avons proposé un procédé de reformage du bruit de quantification inspiré de celui conçu au cours de ma thèse [13, 14]. Considérons le tatouage par STDM d'un signal $x$ par blocs de $N_s$ échantillons. On note $x_{proj}$ le projeté d'un vecteur de $x$ sur un vecteur arbitraire $v$. Le vecteur tatoué par le symbole $m$ s'exprime classiquement :

$$\tilde{x}_m = \underbrace{\big(\mathrm{q}(x_{proj} + d(m)) - d(m)\big)}_{\text{dither-quantification du projeté}} v \quad + \quad \underbrace{(x - x_{proj} v)}_{\text{orthogonal de la projection}} \tag{1.3}$$

où q désigne toujours une quantification, mais pas nécessairement par le quantificateur le plus proche. Soit $\tilde{x}$ le signal tatoué, concaténation des $\tilde{x}_m$ successifs. Posons

$$\nu = \tilde{x} - x \tag{1.4}$$

le bruit de quantification. **Le reformage spectral consiste à imposer à $\nu$ de suivre un modèle ARMA correspondant à la forme du seuil de masquage de $x$. Pour cela, au lieu de quantifier chaque projeté par le quantificateur le plus proche, on cherche la suite des quantificateurs tels que :**

$$\nu(n) = \sigma \sum_{i=0}^{q} b_i w(n-i) - \sum_{i=1}^{p} a_i \nu(n-i) \tag{1.5}$$

$$\text{avec} \quad \sum_{k=-\infty}^{n} w(k)^2 \text{ minimal} \tag{1.6}$$

où $(a_i)_{1 \le i \le p}$, $(b_i)_{1 \le i \le q}$ et $\sigma$ sont respectivement les coefficients AR, les coefficients MA (avec $b_0 = 1$) et le gain du modèle ARMA.

L'algorithme est le suivant. Supposons qu'à un instant donné on ait un ensemble de $N_c$ suites de quantificateurs possibles, que nous appellerons *chemins*. Considérons un ensemble de $K$ quantificateurs possibles pour le nouveau bloc de $N_s$ échantillons, qui sont autant de terminaisons possibles pour chacun des $N_c$ chemins. On a alors $K.N_c$ nouveaux chemins, pour chacun desquels on peut calculer $N_s$ nouveaux échantillons de $\nu$ selon (1.4), puis $N_s$ nouveaux échantillons de $w$ selon (1.5). L'arbre des $K.N_c$ chemins peut être élagué selon le principe de l'algorithme de Viterbi [19]. Dans le cas d'un modèle AR, tous les chemins finissant par les mêmes $\lceil p/N_s \rceil$ derniers quantificateurs ont les mêmes $p$ derniers $\nu(n)$, de sorte que les valeurs suivantes de $w(n)$ seront les mêmes. Par conséquent, parmi ces chemins, on ne garde que celui qui, à ce stade, minimise $\sum_{k=-\infty}^{n} w(k)^2$. Dans le cas d'un modèle ARMA, les chemins

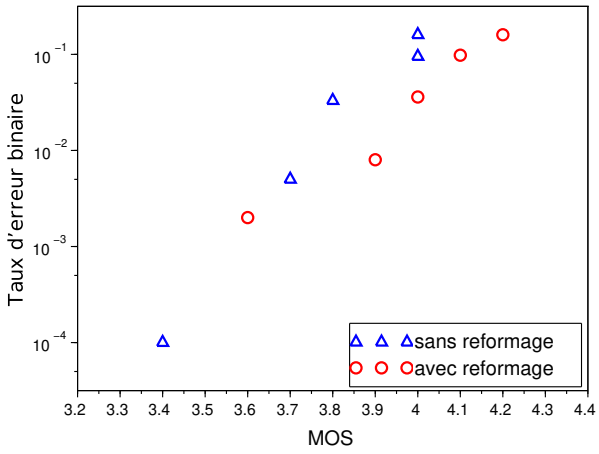



FIGURE 1.2 – Taux d'erreur de détection d'un tatouage STDM à 2 kbit/s en fonction de la qualité du signal tatoué, pour un signal de parole à 8 kHz transmis *via* un canal à bruit blanc gaussien avec un rapport signal à bruit de 40 dB. La qualité est déterminée par le pas de quantification et elle est mesurée par le score MOS fourni par PESQ [9]. Le tatouage peut être considéré comme inaudible à partir d'un MOS de 4.

finissant par les mêmes $\lceil \max(p,q)/N_s \rceil$ derniers quantificateurs n'ont pas nécessairement les mêmes $w(n)$, de sorte qu'il faut comparer sur un passé plus long [1]. Comme dans le décodage d'un code convolutif, on constate empiriquement que tous les chemins conservés convergent, en remontant le temps, vers le même chemin père, ce qui permet de quantifier avec un retard légèrement supérieur à $p.N_s$ dans le cas d'un modèle AR.

Ce travail s'est interrompu en 2009 sans avoir réussi à débugguer les programmes, de sorte que cette méthode n'a pas été publiée. J'ai repris et corrigé les programmes fin 2021, ce qui a donné lieu à des résultats encourageants, illustrés par la figure 1.2 : à un débit élevé (2 kbit/s) le reformage permet d'améliorer le compromis robustesse/distorsion, avec un gain de MOS d'environ 0,2. *A posteriori*, une piste a été négligée, celle d'un choix du sous-espace de projection assurant par lui-même le reformage spectral du bruit de quantification selon le seuil de masquage du signal. Toutefois, elle pose le problème de la connaissance par le récepteur de ce sous-espace. Celle-ci pourrait être fournie *via* le tatouage du bloc précédent, au prix d'une réduction du débit utile de tatouage.

Si l'objectif de court terme de 2007 — améliorer le compromis débit / robustesse / distorsion de la STDM par un reformage du bruit de quantification — n'a été atteint qu'en 2021, ce travail a permis de poursuivre la réflexion sur la manière d'imposer des propriétés au signal dans le domaine fréquentiel par une manipulation dans le domaine temporel. Cette réflexion sera utile par la suite, notamment pour l'égalisation d'histogramme (chapitre III).

1. contrairement à ce que j'ai écrit dans [10] et [13]

## 1.3 Tatouage par modulation des trajectoires des paramètres TVAR

**Collaboration :** Monia Turki, U2S, ENIT

**Encadrements**

- Sondes Maadi, projet de fin d'études d'ingénieur ENIT (2005)
- Imen Samaali, stage de mastère ENIT (2005-2006)
- Mamadou Gueye, stage de M2, Université de Reims (2006)
- Sami Bouzekri, projet de fin d'études d'ingénieur ENIT (2008) et stage de mastère (2008-2009) en co-diplômation ENIT-Paris 5

co-encadré.es avec Monia Turki

**Projet :** WaRRIS (2005-2011)

**Publication :**

[SMA07] I. Samaali, G. Mahé, and M. Turki-Hadj Alouane. Criteria to measure the quality of TVAR estimation for audio signals. In *Proceedings of the 15th European Signal Processing Conference (Eusipco 2007)*, pages 798–802, Poznan, Poland, 2007.

En 2005, Monia Turki, de l'ENIT, a suscité mon intérêt pour la modélisation auto-régressive à coefficients variant dans le temps (time-varying auto-regressive, TVAR) pour les signaux de parole. Nous avons expérimenté les propositions de [18] en co-encadrant le projet de fin d'études ENIT de Sondes Maadi (2005) puis le projet de mastère de Imen Samaali (2005-2006).

Le modèle auto-régressif (AR) de la parole est fondé sur la modélisation du conduit vocal comme une succession de tubes de longueurs et sections différentes. La parole étant non-stationnaire, les coefficients AR sont actualisés par trames, ce qui n'est cependant pas suffisant pour représenter les non-stationnarités, notamment les phonèmes brefs et les transitions entre phonèmes. Le modèle TVAR permet de suivre ces non-stationnarités, en modélisant l'évolution continue de la forme du conduit vocal. Le signal vocal s'exprime :

$$x_t = \sum_{i=1}^{p} a_{i,t} x_{t-i} + \sigma_{e,t} e_t, \quad e_t \sim \mathcal{N}(0,1) \tag{1.7}$$

où les coefficients AR $a_t = [a_{1,t} \dots a_{p,t}]$ et le log de la variance $\sigma^2_{e,t}$ de l'excitation suivent un modèle markovien d'ordre 1.

Ce modèle est utilisé dans [18] pour le débruitage de la parole, que l'on suppose noyée dans un bruit blanc gaussien dont la variance suit aussi un modèle markovien d'ordre 1. On obtient alors une représentation en espace d'état :

$$X_t = A_t X_{t-1} + B_t v_t \quad v_t \sim \mathcal{N}(0,1) \tag{1.8}$$

$$y_t = C_t X_t + D_t w_t \quad w_t \sim \mathcal{N}(0,1) \tag{1.9}$$

où $X_t \triangleq [x_t \dots x_{t-p+1}]^\top$ est l'état du système, $y_t$ l'observation bruitée, et

$$A_t \triangleq \begin{bmatrix} a_t^\top \\ I_{p-1}\, 0_{p-1\times 1} \end{bmatrix}, \; B_t \triangleq \begin{bmatrix} \sigma_{e,t} \\ 0_{p-1\times 1} \end{bmatrix}, \; C_t \triangleq [1\, 0_{1\times p-1}], \; D_t \triangleq [\sigma_{n,t}] \tag{1.10}$$

L'objectif est d'estimer conjointement le modèle $\theta_t \triangleq (a_t, \sigma_{e,t}, \sigma_{n,t})$ et l'état $X_t$.

Dans le cas gaussien linéaire, le signal $x$ peut être estimé de manière optimale par un filtre de Kalman. Dans le cas présent, la représentation n'est gaussienne et linéaire que conditionnellement au modèle, de sorte que la solution ne peut être donnée que par une approximation numérique, notamment par filtrage particulaire.Toutefois, si l'on connaît une approximation de $\mathrm{p}(\theta_t|y_t)$, on peut estimer $\mathrm{p}(x_t|\theta_t, y_t)$ par filtrage de Kalman (le problème étant alors gaussien linéaire) et, partant, $\mathrm{p}(x_t, \theta_t|y_t)$. Cette procédure permet de réduire la variance des estimateurs par rapport à un schéma entièrement par filtrage particulaire.

Nos résultats expérimentaux sont similaires à ceux de [18], à savoir que l'amélioration du rapport signal à bruit (RSB) est modeste (4 dB pour un RSB initial de 0 dB) et qu'elle décroît quand le RSB initial augmente. Il apparaît également que le débruitage est plus performant sur les trames non-voisées que voisées, vraisemblablement du fait de la meilleure adéquation du modèle de l'excitation.

**Nous nous sommes plus particulièrement intéressé à la mesure de la qualité de l'estimation TVAR.** Considérons $Y_1 \dots Y_t$ les variables aléatoires associées aux observations $y_1 \dots y_t$, $u_t = \Pr(Y_t \leq y_t|y_{1:t-1})$ et $v_t = \Phi^{-1}(u_t)$, avec $\Phi$ la fonction de répartition normale. Selon [18], le modèle TVAR est valide si $v_t$ est indépendante et identiquement distribuée selon la loi normale. Nous avons proposé une approche différente. En supposant que le signal est produit par un système TVAR de même ordre que le modèle, la comparaison entre les coefficients originaux et leurs estimées peut être agrégée dans une mesure globale et perceptivement pertinente, la distance cepstrale (les coefficients cepstraux s'exprimant aisément en fonction des coefficients AR). Dans le cas où le modèle original n'est pas connu, notre proposition est que **le modèle estimé est valide s'il existe un bruit blanc gaussien stationnaire qui a pu produire le signal estimé $\hat{x}$ par excitation du modèle TVAR estimé.** Soit $\hat{e}$ le signal obtenu par inversion du modèle estimé :

$$\hat{e}(t) = \frac{1}{\hat{\sigma}_{e,t}} \left( \hat{x}(t) - \sum_{i=1}^{p} \hat{a}_{i,t} \hat{x}(t-i) \right) \tag{1.11}$$

Le modèle est valide si $\hat{e}$ est stationnaire, blanc et gaussien. Nous avons validé notre proposition en montrant que des indices de stationnarité, de blancheur et de gaussianité sont fortement corrélés à la distance cepstrale [SMA07]. Cette approche à l'avantage d'avoir une plus faible complexité que celle de [18]

Le projet WaRRIS devenant la priorité en 2006, nous avons cherché à y intégrer ces travaux. Nous les avons repris en 2008 et 2009, dans le cadre du projet de fin d'études ENIT de Sami Bouzekri puis de son projet de mastère, toujours co-encadré par Monia Turki. Il s'agissait d'exploiter notre début de maîtrise de l'estimation de modèles TVAR par filtrage particulaire pour concevoir un **tatouage fondé sur la modulation des paramètres TVAR**. Cette idée est motivée par ce que le fondement physiologique de cette représentation lui permet de capturer l'essence perceptive d'un signal de parole (contrairement à des échantillons temporels ou des coefficients temps-fréquence) et d'être ainsi robuste à un canal modérément dégradé (par exemple une compression à bas-débit).

La représentation d'état précédente, adaptée au débruitage, n'est plus pertinente pour cette application. D'autre part, le modèle markovien d'ordre 1 des coefficients TVAR est sans doute mathématiquement efficace, mais il ne reflète pas les variations

douces de la forme du conduit vocal. Nous l'avons donc remplacé par un modèle markovien d'ordre 1 de l'accélération des modules et des arguments des pôles du modèle TVAR, plus représentatif de l'évolution temporelle des résonances du conduit vocal. Il en résulte la représentation d'état suivante pour un modèle TVAR d'ordre 2 avec 2 pôles complexes conjugués [2] de module $\rho_t$ et d'arguments $\pm 2\pi\nu_t$ :

$$\begin{bmatrix} Z_t^\rho \\ Z_t^\nu \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} Z_{t-1}^\rho \\ Z_{t-1}^\nu \end{bmatrix} + \begin{bmatrix} B_\rho & 0 \\ 0 & B_\nu \end{bmatrix} \begin{bmatrix} w_t^\rho \\ w_t^\nu \end{bmatrix} + \begin{bmatrix} m_\rho \\ m_\nu \end{bmatrix} \tag{1.12}$$

$$\text{avec} \quad Z_t^\rho = \begin{bmatrix} \rho_t \\ \rho_t' \\ \rho_t'' \end{bmatrix}, \quad Z_t^\nu = \begin{bmatrix} \nu_t \\ \nu_t' \\ \nu_t'' \end{bmatrix}, \quad A = \begin{bmatrix} \mu & \frac{1+\mu}{2} & \frac{1+\mu}{4} \\ 0 & \mu & \frac{1+\mu}{2} \\ 0 & 0 & \mu \end{bmatrix},$$

$$B_\rho = \begin{bmatrix} 1/4 \\ 1/2 \\ 1 \end{bmatrix} \sigma_\rho, \quad B_\nu = \begin{bmatrix} 1/4 \\ 1/2 \\ 1 \end{bmatrix} \sigma_\nu, w_t^\rho \sim \mathcal{N}(0,1),\ w_t^\nu \sim \mathcal{N}(0,1)$$

$$m_\rho = \begin{bmatrix} (1-\mu)\overline{\rho} \\ 0 \\ 0 \end{bmatrix}, \quad m_\nu = \begin{bmatrix} (1-\mu)\overline{\nu} \\ 0 \\ 0 \end{bmatrix}$$

$$x_t = 2\rho_t \cos(2\pi\nu_t) x_{t-1} - \rho_t^2 x_{t-2} + \sigma_{e,t} e_t \tag{1.13}$$

La méthode précédente d'estimation par filtrage particulaire combiné à un filtrage de Kalman est utilisée, avec cependant des différences notables : d'une part, les paramètres à estimer par filtrage particulaire sont réduits à $\theta_t = \{\sigma_{e,t}\}$, ce qui permet de réduire le nombre de particules nécessaires ; d'autre part, l'équation d'observation (1.13) n'étant pas linéaire, on utilise un filtre de Kalman étendu.

Le tatouage proposé consiste à ajouter un signal NRZ ou RZ bipolaire aux trajectoires des modules et des fréquences des pôles. L'amplitude de ce signal est ajustée de manière à garantir l'inaudibilité du tatouage. Le signal tatoué suit alors la représentation d'état (1.12,1.13) en remplaçant $\overline{\rho}$ par $\overline{\rho} + \lambda_t$ et $\overline{\nu}$ par $\overline{\nu} + \lambda_t$, avec $\lambda_t$ constant par morceaux, dépendant du symbole inséré et de la modulation choisie (NRZ ou RZ).

La première étape de l'extraction du tatouage consiste à estimer les paramètres, avec ici $\theta_t = \{\sigma_{e,t}, \lambda_t\}$ estimé par filtrage particulaire. Puis, selon le type de modulation, le symbole inséré dans un bloc de signal peut être estimé de deux manières : dans le cas d'une modulation NRZ, l'estimée est la moyenne temporelle de l'estimation de $\lambda_t$ sur une durée symbole ; dans le cas d'une modulation RZ, nous estimons le symbole inséré par détection des discontinuités dans la trajectoire estimée de la fréquence du pôle, à l'aide du détecteur algébrique présenté dans [6, 16].

Les simulations ont été réalisées sur des signaux synthétiques TVAR d'ordre 2, avec une modulation binaire et un débit binaire de tatouage de 100 bit/s pour une fréquence d'échantillonnage de 16 kHz. Le taux d'erreur de détection du tatouage est de 2% avec la première méthode, 5% avec la seconde. Le tatouage semble relativement robuste au bruit jusqu'à un RSB de 20 dB (taux d'erreur binaire de 10%) et à certaines compressions audio (taux d'erreur de 12% avec un codage AAC à 8kbit/s, mais de 35% avec un codage GSM).

2. L'étude sur ce modèle simple est un préliminaire avant l'application à la parole, qui suppose des modèles d'ordre 10 environ

Nous avons interrompu en 2011 cette étude (non publiée), qui nécessitait un travail important de consolidation et d'extension à des signaux de parole réels, pour un espoir limité d'atteindre le haut-débit de tatouage souhaité pour nos applications. Le détecteur algébrique de discontinuités étudié au cours de ce travail a cependant été utile pour la suite : c'est cet outil que nous avons employé dans la thèse de Imen Samaali pour détecter les attaques musicales (voir section 1.4 et chapitre II).

## 1.4 Tatouage par étalement de spectre

**Collaborations :**

— Monia Turki, U2S, ENIT

— Sonia Larbi, U2S, ENIT

**Encadrement :** Imen Samaali, thèse en co-tutelle Paris 5 - ENIT (2006-2012), co-encadrée avec Monia Turki

**Projets :** WaRRIS et EReQCA (2006-2012)

**Publications :**

[SMT12] Imen Samaali, Gaël Mahé, and Monia Turki. Watermark-aided pre-echo reduction in low bit-rate audio coding. *J. Audio Eng. Soc*, 60(6) :431–443, june 2012.

Au moment où nous souhaitions exploiter le tatouage à des fins de correction du signal hôte, aucune des techniques précédemment évoquées n'offrait un compromis inaudibilité-débit-robustesse satisfaisant. Nous nous sommes donc tourné vers une technique limitée en débit mais largement éprouvée au sein de nos équipes, à savoir le tatouage par étalement de spectre [11]. Celui-ci avait fait l'objet des thèses de Leandro Gomes au LIPADE [5] et de Sonia Larbi à l'U2S [12], en lien avec celle de Cléo Baras à l'ENST [2].

Le principe est d'ajouter au signal hôte une série de bruits codant chacun un symbole, tel que le spectre du bruit soit sous le seuil de masquage fréquentiel du signal, comme illustré par la figure 1.3. Le modulateur (*Mod*) associe à chaque symbole $a_k$ un vecteur de bruit blanc gaussien, fourni par le dictionnaire $D$. La succession de ces vecteurs constitue le signal modulé $v(t)$.Celui-ci est ajouté au signal audio $x(t)$ après reformage spectral par le filtre $H_t(f)$ assurant que le spectre de sa sortie $w(t)$ soit sous le seuil de masquage de $x(t)$, déterminé selon le modèle psychoacoustique (MPA) du codeur MPEG-1 [8]. Le canal de communication peut correspondre ici à diverses altérations : compression audio, bruit, transmission dégradée... Du point de vue de la théorie des communications numériques, le récepteur détecte les symboles émis dans un contexte difficile : un « bruit » (le signal audio) très fort, corrélé et non-gaussien ; des interférences entre symboles produites par le filtre $H$ ; un canal non-stationnaire du fait de la variabilité au cours du temps de $H$ ; des dégradations additionnelles liées au canal de communication. Nous avons retenu la solution proposée par [12], qui repose sur un égaliseur par zero-forcing et un filtre de Wiener. L'inversion du filtre de mise en forme $H$, inconnu en réception, par l'égaliseur $A$ repose sur l'hypothèse que le seuil de masquage de $\hat{y}(t)$ est très proche de celui de $x(t)$. Le filtre de Wiener $G$ est calculé à partir des fonctions d'autocorrélation de

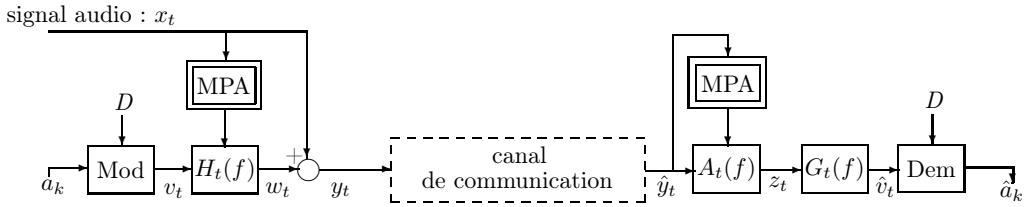



FIGURE 1.3 – Chaîne de tatouage par étalement de spectre.

$z(t)$ et de $v(t)$, cette dernière pouvant être obtenue par la simple connaissance du dictionnaire de modulation $D$ en réception.

Dans le cadre de la thèse de Imen Samaali, nous avons prolongé l'étude menée par Sonia Larbi et Cléo Baras sur la robustesse de ce tatouage à la compression [SMT12], en considérant les formats MP3, AAC et AAC+, à divers débits offrant une qualité entre acceptable et transparente. Ces compressions éliminent ou quantifient grossièrement les composantes haute-fréquence et synthétisent de manière approximative celles de fréquences intermédiaires, par des procédés d'extension de bande. De ce fait, l'hypothèse utilisée pour l'égalisation, selon laquelle le signal reçu a le même seuil de masquage que le signal original, n'est valable que pour des fréquences inférieures à 5 kHz, voire 3,5 kHz pour AAC+. Le filtrage passe-bas (à ces fréquences de coupure) de $v(t)$ avant le filtre de mise en forme $H$ permet d'améliorer la robustesse du tatouage : la compression n'augmente le taux d'erreur que d'un facteur inférieur à 10, contre 100 à 1000 sans filtrage passe-bas.

Nous avons aussi mis en évidence la faible robustesse du tatouage à la compression dans le cas particulier de signaux percussifs, liée à la mauvaise reproduction des attaques par les codecs audio (phénomène de pré-écho). Notre système de tatouage intègre un détecteur algébrique de discontinuités issu de [6, 16] (qui reste efficace en réception malgré l'amolissement des attaques par le codec) et les trames à attaques ne sont pas tatouées. À débit global de tatouage constant, le taux d'erreur de détection est alors divisé par 100 pour les instruments percussifs testés. Enfin, nous avons montré expérimentalement que le système proposé est robuste aux compressions multiples : le taux d'erreur augmente très faiblement au fil des codage/décodage.

Le tatouage par étalement de spectre n'a pas été considéré une technique indépendante de son contexte d'utilisation. Son étude a été liée à ce contexte en ce que les altérations du signal hôte auquel il doit être robuste sont celles que l'on va chercher à corriger grâce à l'information insérée. Une partie du travail est ainsi déjà faite pour cette correction, qui sera présentée au chapitre II.

## 1.5 Conclusion

Nous avons étudié le tatouage audio d'abord pour lui-même, comme un jeu gratuit, sans autre objectif que de maximiser le débit d'insertion sous les contraintes de robustesse et d'inaudibilité. Puis la construction du projet WaRRIS et l'engagement dans celui-ci ont modifié notre approche : le tatouage devenait un maillon d'une chaîne de traitement audio, auquel était assigné un certain niveau de performance en termes de compromis débit-robustesse-inaudibilité.

Dans cette perspective, nous avons conservé le tatouage par étalement de spectre, qui offrait les meilleures performances, et nous nous sommes concentré sur l'exploitation du tatouage par diverses applications. Le bilan des travaux présentés dans ce chapitre ne saurait cependant se résumer à la sélection d'une technique de tatouage pour un maillon d'une chaîne de traitement : au delà des performances des méthodes de tatouage étudiées, une partie du travail d'élaboration a fourni des outils théoriques et pratiques aux travaux présentés par la suite.

Le chapitre suivant montre comment un tatouage à débit modéré (de l'ordre d'une centaine de bit/s) et robuste à la compression peut enrichir utilement un signal audio, en fournissant au récepteur des informations facilitant la correction du signal. Finalement, le développement, dans le cadre du projet WaRRIS, d'une conception étendue de la notion de tatouage fera disparaître la notion même de débit de tatouage, le tatouage prenant les fonctions de modifier les propriétés du signal hôte (chapitre III) ou d'enregistrer les distorsions subies par celui-ci (chapitre IV).

## 1.6 Références bibliographiques


[1] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J.-M. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J.-P. Léoni, J. Liénard, N. Moreau, and P. Nguyen. ARTUS : synthesis and audiovisual watermarking of the movements of a virtual agent interpreting subtitling using cued speech for deaf televiewers. *Modelling, measurement and control C*, 67SH(2, supplement : handicap) :177–187, 2006. AMSE - ISSN : 1259-5977.

[2] C. Baras. *Tatouage informé de signaux audio numériques.* Thèses, Télécom ParisTech, Dec. 2005.

[3] B. Chen and G. W. Wornell. Quantization index modulation : A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. on Information Theory*, 47(4) :1423–1443, 1999.

[4] T. Ciloglu and S. U. Karaaslan. An improved all-pass watermarking scheme for speech and audio. In *Proceedings of IEEE International Conference on Multimedia and Expo(ICME)*, 2000.

[5] L. De Campos Teixeira Gomes. *Tatouage de signaux audio.* Thèses, Université René Descartes - Paris V, 2002.

[6] M. Fliess, C. Join, and M. Mboup. Algebraic change-point detection. *Applicable Algebra in Engineering, Communication and Computing*, 21(2) :131–143, 2010.

[7] D. Gruhl, A. Lu, and W. Bender. Echo hiding. In R. Anderson, editor, *Workshop on Information Hiding*, pages 295–315, Berlin, Heidelberg, 1996. Springer.

[8] ISO/IEC. *Norm 11172-3 : Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3 : Audio*, 1993.

[9] ITU-T. Recommendation P.862 : Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[10] N. S. Jayant and P. Noll. *Digital Coding of Waveforms, Principles and Applications to Speech and Video.* Prentice Hall, 1984.

[11] D. Kirovski and H. Malvar. Spread-spectrum watermarking of audio signals. *IEEE Transactions on Signal Processing*, 51(4) :1020–1033, 2003.

[12] S. Larbi. *Structures d'égalisation en tatouage audio numérique*. Thèses, Télécom ParisTech et École Nationale d'Ingénieurs de Tunis, 2005.

[13] G. Mahé and A. Gilloire. Quantization noise spectral shaping in instantaneous coding of spectrally unbalanced speech signals. In *Proc. IEEE Workshop on Speech Coding*, pages 56–58, Tsukuba, Ibaraki, Japon, October 2002.

[14] G. Mahé and A. Gilloire. Contrôle de l'audibilité du bruit de quantification induit par la pré-distorsion d'un signal de parole. In *Proc. GRETSI*, pages 237–240, Paris, France, September 2003.

[15] H. M. A. Malik, R. Ansari, and A. A. Khokhar. Robust data hiding in audio using allpass filters. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4) :1296–1304, 2007.

[16] M. Mboup, C. Join, and M. Fliess. A delay estimation approach to change-point detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[17] C. Podilchuk and E. Delp. Digital watermarking : algorithms and applications. *IEEE Signal Processing Magazine*, 18(4) :33–46, 2001.

[18] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill. Particle methods for bayesian modeling and enhancement of speech signals. *IEEE Transactions on Speech and Audio Processing*, 10(3) :173–185, 2002.

[19] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269, 1967.

[20] Y. Yardimci, A. E. Cetin, and R. Ansari. Data hiding in speech using phase coding. In *Proceedings of Eurospeech 97*, volume 3, pages 1679–1682, 1997.

[21] E. Zwicker and R. Feldtkeller. *The ear as a communication receiver*. Acoustical Society of America, New York, 1999.

# Chapitre 2

# Du bruit pour informer

## Le tatouage réflexif

**Sommaire**



Dans le cadre du projet WaRRIS, nous avons proposé une nouvelle approche du tatouage pour l'enrichissement de contenu, le *tatouage d'accentuation* ou *tatouage réflexif* : il s'agit de **tatouer le signal par des informations sur lui-même, de manière à faciliter sa correction à la sortie d'un canal de communication dégradé**, comme illustré par la figure 2.1. En d'autres termes, on bruite le signal pour informer un traitement ultérieur.

Cette approche a été explorée par quelques rares travaux à la même époque (autour de 2010), pour corriger des pertes de blocs ou étendre la largeur de bande, en téléphonie. Une idée simple est de tatouer le signal par une version compressée de lui-même [26], ce qui nécessite cependant des débits de tatouage peu compatibles avec les contraintes réelles du canal [1]. De manière plus réaliste, Geiser *et al.* [4] n'insèrent qu'une information servant à aider, en réception, des méthodes classiques d'interpolation entre blocs, adaptées à un codage spécifique de la parole (AMR wideband) : il en résulte un débit total de tatouage de 2 kbit/s, incluant un codage correcteur d'erreur. Le gain en qualité sur un canal avec perte de paquets reste modeste, inférieur à 0,3 en échelle MOS. Le même principe est utilisé par [25] pour étendre la bande téléphonique. La bande haute (3400-8000 Hz) est reconstruite à partir de la bande basse (forte corrélation) et d'une information auxiliaire, transmise par tatouage, contenant notamment les paramètres de l'enveloppe spectrale dans les hautes

1. Les auteurs cités présentent des résultats peu réalistes : tatouage très basique par modulation des 2 bits de poids faible, MOS de 4,2, débit de tatouage de 16 kbit/s, effacement d'un tiers du signal, MOS après correction de 3,6... mais en considérant un canal transparent.

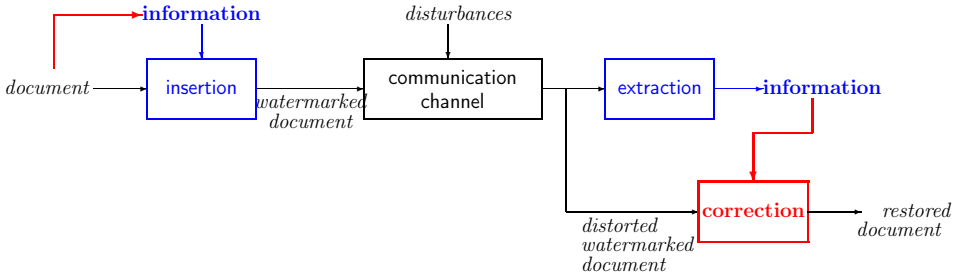



FIGURE 2.1 – Schéma de principe du tatouage réflexif.

fréquences. Cette information nécessite un débit de tatouage de 4 kbit/s, robuste à un canal téléphonique classique (dégradations modérées).

Le tatouage réflexif modifie l'équilibre des contraintes débit / robustesse / inaudibilité évoqué dans le chapitre précédent. Les applications visées nécessitent un haut débit d'information, que ne permettent pas la plupart des méthodes de tatouage audio disponibles. Comme l'objectif est de corriger les dégradations d'un canal de transmission, le tatouage doit évidemment être robuste à ces dégradations. Il est toutefois possible de concentrer cette contrainte de robustesse sur ces dégradations spécifiques tout en la relâchant sur les autres dégradations (d'autant plus que les « attaques malicieuses », préoccupation du tatouage sécuritaire, sont hors-sujet ici). Enfin, la contrainte d'inaudibilité du tatouage, stricte dans les applications usuelles, peut être relâchée, l'objectif étant que le son après correction soit de meilleure qualité que le son non-tatoué en sortie du canal.

Un principe à retenir des exemples présentés est qu'il est plus économique, en termes de débit de tatouage, d'insérer une information auxiliaire permettant d'aider des méthodes classiques de correction à partir du signal seul que d'insérer toutes les données manquantes.

Ce chapitre présente l'application du principe de tatouage réflexif dans deux contextes : dans les sections 2.1 et 2.2, le canal de communication de la figure 2.1 consiste en un codage audio à bas-débit suivi d'un décodage (avec éventuellement une transmission entre les deux) et il s'agit de corriger certaines dégradations résultant de la compression ; dans la section 2.9, le signal est multi-pistes, le canal consiste en un mixage stéréo suivi d'une transmission et il s'agit de démixer le mélange stéréo (séparation de sources).

## 2.1 Codage audio : suppression de pré-écho assistée par tatouage

**Co-encadrement :** thèse de Imen Samaali, en co-tutelle Paris 5 - ENIT (2006-2012), co-encadrée avec Monia Turki (U2S, ENIT)

**Projets :** WaRRIS et EReQCA (2006-2012)

**Publications :**

[SAM09] I. Samaali, M. Turki-Hadj Alouane, and G. Mahé. Temporal envelope correction for attack restoration in low bit-rate audio coding. In *Proceedings*

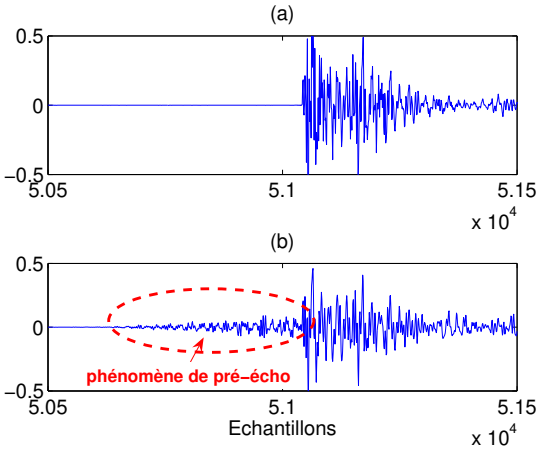



FIGURE 2.2 – Exemple de pré-écho sur un signal de castagnettes : (a) signal original, (b) signal synthétisé par un codeur MP3 à 48 kbit/s.

*of the 17th European Signal Processing Conference (Eusipco 2009)*, pages 929–933, Glasgow, United Kingdom, August 2009.

[SAM10] I. Samaali, M. Turki-Hadj Alouane, and G. Mahé. Attack localization based on algebra detector for pre-echo reduction in low bit-rate audio coding. In *Proc. 5th International Symposium on Image/Video Communications and Mobile Networks (ISIVC)*, Rabat, Morocco, September 2010.

[SMT12] Imen Samaali, Gaël Mahé, and Monia Turki. Watermark-aided pre-echo reduction in low bit-rate audio coding. *J. Audio Eng. Soc*, 60(6) :431–443, June 2012.

L'objet de la thèse de Imen Samaali était de démontrer la validité du principe du tatouage réflexif présenté en introduction de ce chapitre, dans le contexte particulier de la compression audio à bas débit [2] (donc avec pertes et légère dégradation de qualité audio) : **il s'agit de montrer qu'il est possible de restaurer (au moins partiellement) la qualité d'un signal audio à l'issue d'une compression-décompression, au moyen d'informations insérées par tatouage dans le signal avant compression**. Ici, le canal de communication de la figure 2.1 consiste en un codage-décodage.

Lorsque le taux de compression est trop fort, les codeurs perceptifs implémentés selon les normes MPEG produisent sur les signaux audio percussifs un *pré-écho* juste avant les attaques, comme illustré sur la figure 2.2. Si celui-ci est suffisamment bref et faible, il peut être masqué par l'attaque (phénomène de masquage antérieur [30]), sinon il est perçu comme un bruit d'aspiration avant l'attaque. Ce pré-écho résulte du codage par transformée sur des fenêtres de longueur fixe, qui produit un bruit de quantification spectralement reformé pour être sous le seuil de masquage fréquentiel, mais d'énergie répartie uniformément sur la durée de la fenêtre. Cette énergie est alors déterminée par celle de l'attaque, de sorte qu'elle dépasse nettement celle du signal avant l'attaque.

2. Nous considérons ici comme bas-débit un débit inférieur au débit minimal assurant une qualité transparente (dégradation inaudible), soit 96 kbit/s pour MP3 et 64 kbit/s pour AAC, pour un signal mono.

Ce pré-écho peut être réduit par des options spécifiques des codeurs. Dans les codeurs MP3, l'option *temporal masking* (TM) adapte la taille des fenêtres au signal [19] : des fenêtres longues (1024 échantillons) sont utilisées pour les segments stationnaires, des fenêtres courtes (64 échantillons) pour les segments non-stationnaires, ce qui permet de réduire la durée maximale des pré-échos et, partant, de les masquer. Le codeur AAC dispose d'une option *temporal noise shaping* (TNS [7]) permettant de reformer l'enveloppe temporelle du bruit de quantification selon celle du signal. Le principe est le suivant : de même qu'une quantification scalaire prédictive en boucle ouverte d'un signal dans le domaine temporel se traduit par une erreur de quantification dont la densité spectrale de puissance est de même forme que celle du signal, quantifier le résidu de prédiction linéaire de la représentation fréquentielle produit un bruit dont la répartition temporelle de la puissance suit celle du signal. En pratique, pour des signaux très percussifs, la réduction du pré-écho opérée par les options TM et TNS est limitée, notamment à faible débit.

**Pour corriger le pré-écho résiduel, nous avons proposé de tatouer le signal avant codage par une représentation compacte de l'enveloppe temporelle de chaque fenêtre et, après décodage et extraction du tatouage, d'aligner l'enveloppe temporelle du signal décodé sur l'enveloppe originelle.**

Nous adoptons la modélisation de l'enveloppe temporelle par prédiction linéaire dans le domaine fréquentiel (FDLP, [15]), qui repose sur des principes similaires à ceux du TNS. De même que l'enveloppe spectrale du signal peut être approchée par un modèle auto-régressif (AR) dont les coefficients sont fournis par l'autocorrélation du signal, l'enveloppe temporelle peut être approchée par un modèle AR dont les coefficients sont fournis par l'autocorrélation de la partie non-orthogonale de la transformée en cosinus discret impaire de type I du signal.

Nous remplaçons le modèle AR par un modèle ARMA, qui nécessite moins de coefficients pour une même finesse d'approximation. Les coefficients AR et MA sont convertis en coefficients LSF (line spectral frequencies), plus appropriés à la quantification, et sont codés selon une quantification vectorielle, dont le dictionnaire est appris selon l'algorithme de Linde-Buzo-Gray [12] sur un corpus d'apprentissage composé de musiques variées. Cependant, dans le cas de signaux percussifs, même avec un ordre de modélisation élevé, il est difficile de suivre les attaques, puisque le modèle ARMA est une version lissée de l'enveloppe temporelle. C'est pourquoi nous proposons de couper chaque fenêtre transitoire à l'instant de l'attaque et de représenter chacune des deux sous-fenêtres par un modèle propre, ce qui nécessite de détecter et localiser les attaques. Les fenêtres sans attaque sont divisées en deux sous-fenêtres de même durée.

Pour la détection des attaques, nous utilisons l'algorithme du codeur AAC [1], qui découpe la fenêtre en 16 sous-fenêtres et détecte une transitoire si l'énergie d'une des sous-fenêtres dépasse d'un certain facteur l'énergie moyenne des sous-fenêtres voisines. Nous avons considéré deux méthodes de localisation des attaques. La première est un indice de stationarité proposé par [24], qui mesure, à chaque instant, une distance entre les représentations temps-fréquence du signal juste avant et juste après l'instant. La seconde est le détecteur algébrique de discontinuités [3, 16]. Celui-ci présente deux avantages : d'une part il a une complexité en $N^2$ pour une fenêtre de longueur $N$, contre $N^2 \log N$ pour l'indice de stationarité ; d'autre part, comme son calcul implique des intégrations qui réduisent l'effet du bruit, il est beaucoup plus

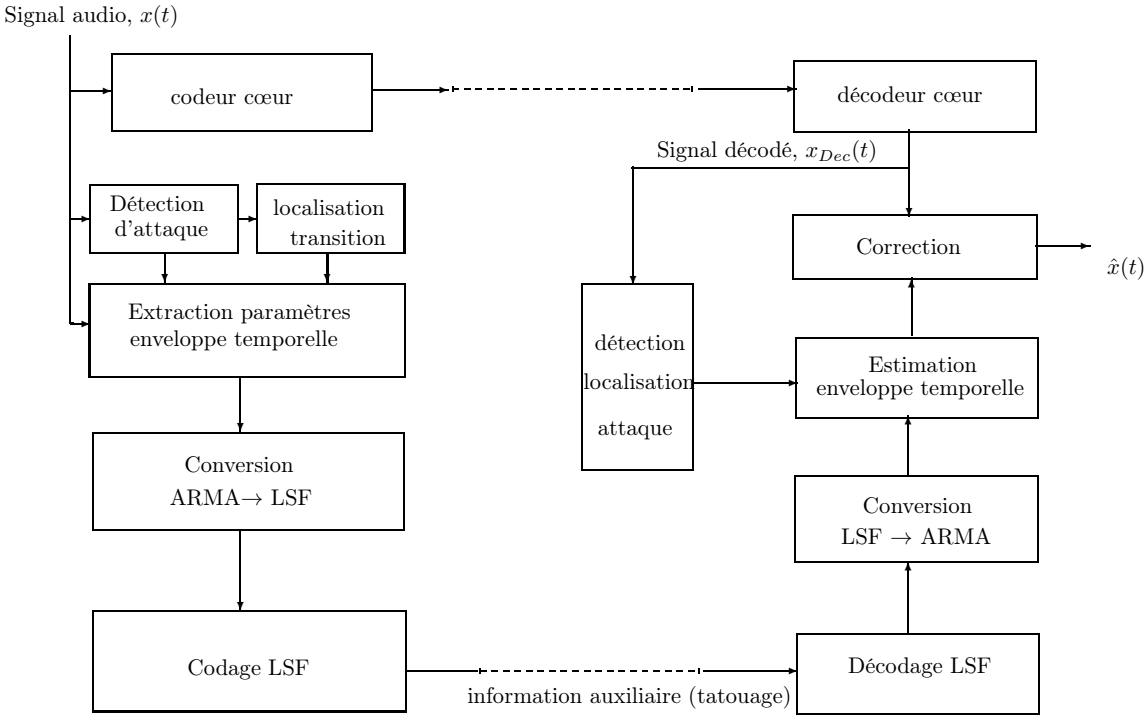



FIGURE 2.3 – Structure complète du système de réduction de pré-écho.

robuste au codage. Ainsi, alors que les attaques sont nettement moins nettes dans le signal décodé, le détecteur algébrique reste efficace pour les localiser précisément, ce qui évite de transmettre l'instant de transition.

Le système global de correction est schématisé par la figure 2.3. Ici, on considère dans un premier temps la transmission distincte de deux flux binaires : l'un contenant le signal audio codé, l'autre les informations auxiliaires permettant la correction. Celles-ci sont constituées des coefficients LSF représentant les coefficients ARMA $\{(a_i)_{1\geq i\geq p}; (b_i)_{1\geq j\geq q}\}$ de chaque sous-fenêtre, où les $b_i$ sont normalisés par $b_0$. Après décodage, les attaques sont détectées et localisées comme lors du codage, chaque fenêtre est découpée selon l'existence et, le cas échéant, la position de l'attaque, les coefficients ARMA originels sont décodés. La même modélisation FDLP est appliquée à l'enveloppe de chaque sous-fenêtre du signal décodé. On note $\{(\tilde{a})_{1\geq i\geq p}; (\tilde{b})_{0\geq j\geq q}\}$ les coefficients ARMA correspondants. Les enveloppes originelles $e(t)$ et dégradée $\tilde{e}(t)$ s'expriment :

$$e(t) = \left|H\left(\mathrm{e}^{\mathrm{j}t}\right)\right|, \quad \text{avec } H(z) = \tilde{b}_0 \frac{1+\sum_{i=1}^{q} b_i z^{-i}}{1+\sum_{i=1}^{p} a_i z^{-i}} \tag{2.1}$$

$$\tilde{e}(t) = \left|\tilde{H}\left(\mathrm{e}^{\mathrm{j}t}\right)\right|, \quad \text{avec } \tilde{H}(z) = \frac{\tilde{b}_0+\sum_{i=1}^{q} \tilde{b} z^{-i}}{1+\sum_{i=1}^{p} \tilde{a} z^{-i}} \tag{2.2}$$

La correction d'enveloppe consiste simplement à multiplier le signal décodé par $e(t)/\tilde{e}(t)$

Les performances ont d'abord été évaluées en considérant la transmission des informations de correction *via* un canal auxiliaire sans erreur. L'utilisation d'un modèle ARMA 5,3 permet une modélisation suffisamment précise des enveloppes temporelles pour un débit d'information auxiliaire d'environ 400 bit/s. La qualité du signal corrigé est mesurée par le score ODG (objective difference grade) fourni par PEMO-Q

(PErceptual MOdel of Quality assessment [9]), qui prédit la dégradation subjective qui serait mesurée par une évaluation selon [10]. Selon cette mesure, l'amélioration de la qualité est notable pour des signaux très percussifs ayant subi une compression MP3 : par exemple, pour un enregistrement de castagnettes, la correction du pré-écho permet d'atteindre une qualité considérée comme transparente (ODG $\geq 1$) à partir d'un débit de codage de 56 kbit/s, alors que sans correction il faut, en plus de l'option TM, un débit supérieur à 96 kbit/s. L'amélioration est moindre dans le cas d'un codage AAC : dans l'exemple précédent, le signal codé-décodé à 64 kbit/s avec TNS (qualité transparente) se compare au signal codé-décodé à 56 kbit/s avec correction. Enfin, le système proposé s'avère robuste aux compressions en cascade : alors que sans correction, l'ODG décroît rapidement au fil des codage-décodage, il reste stable avec notre correction.

Dans un second temps, nous avons remplacé le canal auxiliaire par un tatouage audio par étalement de spectre (voir 1.4) et étudié l'influence de la détection du tatouage sur les performances du réducteur de pré-écho. La difficulté est que la correction est d'autant plus nécessaire que le débit du codec est faible et que plus ce débit est faible, plus le tatouage utile à la correction est dégradé par la compression. Pour déterminer les conditions de fonctionnement du système, nous avons tracé d'une part les courbes d'ODG du réducteur de pré-écho en fonction du taux d'erreur binaire (TEB) sur le canal auxiliaire, d'autre part les courbes de TEB de la détection du tatouage en fonction du débit d'insertion, pour différents débits de codecs MP3 et AAC. Il en résulte que quel que soit le débit du codec, l'ODG reste maximal tant que le TEB est inférieur à $10^{-2}$, ce qui nécessite de ne pas dépasser un débit de tatouage de 150 bit/s, voire 100 bit/s pour résister à des codage-décodage multiples. Or la correction présentée *supra* suppose un débit d'information auxiliaire de 400 à 500 bit/s.

Nous avons donc adapté la correction à la contrainte de débit en ne traitant que les trames à attaque et leurs prédécesseuses, le pré-écho d'une trame pouvant d'étendre sur la trame précédente[3]. Cette adaptation permet de conserver les performances du système, illustrées par la figure 2.4, avec un débit de tatouage de 50 bit/s. Pour le codeur AAC, l'amélioration de la qualité est plus modeste (voir [SMT12]), ce qui peut s'expliquer par l'utilisation de trames plus courtes pour les transitoires, qui réduisent la durée du pré-écho.

**Nous avons ainsi montré que le tatouage réflexif permet d'améliorer nettement la qualité audio des signaux percussifs dégradés par une compression MPEG à faible débit.**

Un exemple de correction est présenté sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/HDR/`.

3. Le tatouage étant réparti sur tout le signal sauf les trames à attaque, il peut se poser un problème de synchronisation, que nous n'avons pas traité.

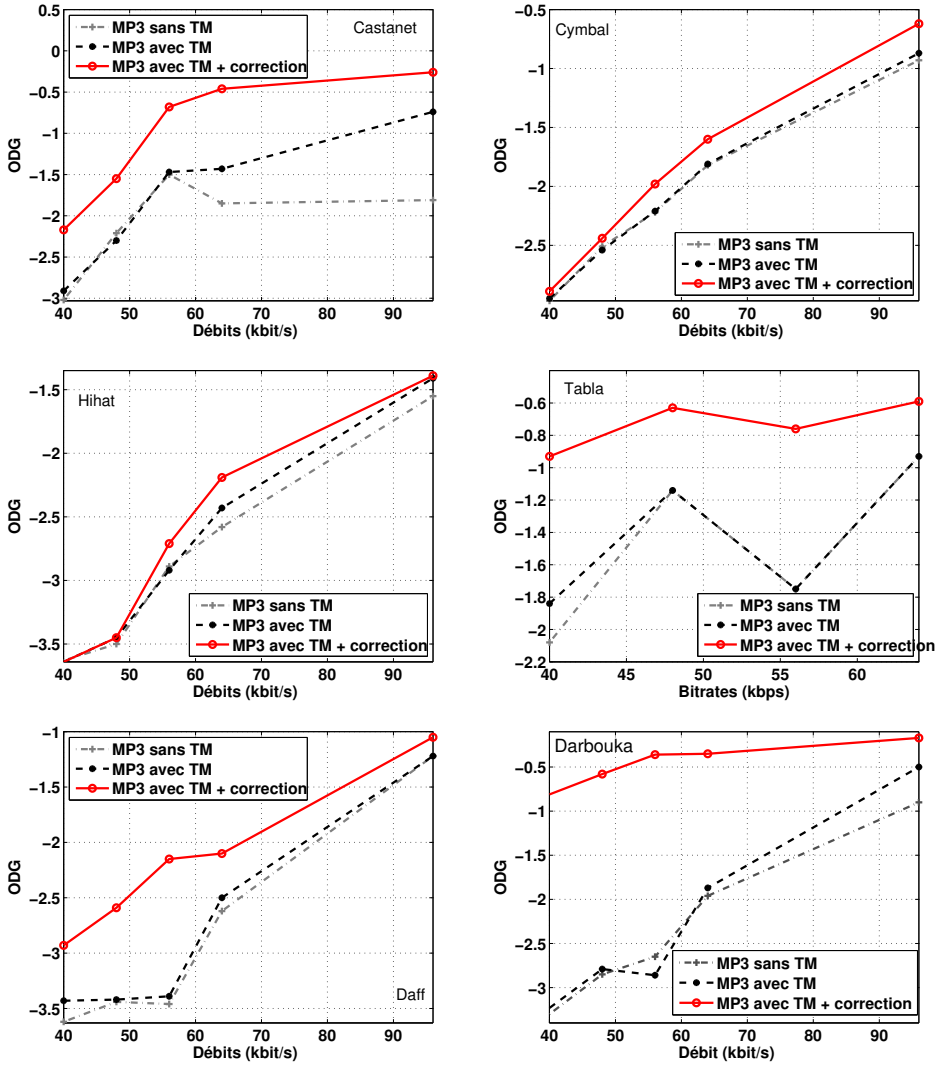



FIGURE 2.4 – Pour six signaux tests (instruments percussifs), évaluation des performances du système de réduction de pré-écho dans le cas de la compression MP3 : qualité prédite (ODG) en fonction du débit du codeur, avec et sans correction.

## 2.2 Codage audio : restauration de tonales assistée par tatouage

**Co-encadrement :** thèse de Imen Samaali, en co-tutelle Paris 5 - ENIT (2006-2012), co-encadrée avec Monia Turki (U2S, ENIT)

**Projets :** WaRRIS et EReQCA (2006-2012)

**Publication :**

[SMA15] I. Samaali, G. Mahé, and M. T. H. Alouane. High-frequency tonal components restoration in low-bitrate audio coding using multiple spectral translations. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1053–1057, Aug 2015.

Dans la thèse de Imen Samaali, nous nous sommes intéressé à un autre défaut des codeurs audio à bas-débit, la synthèse erronée des composantes hautes fréquences dans le mécanisme de réplication de bande spectrale (*spectral band replication*, SBR) utilisé par le MP3pro et le HE-AAC (*High-Efficiency Advanced Audio Coding*, également appelé AAC+). Ces codeurs reposent respectivement sur le codeur MP3 et sur le codeur AAC pour les composantes en deçà d'une certaine fréquence de coupure (d'autant plus basse que le débit est faible) et codent les hautes fréquences par une information minimale (de l'ordre de 2 kbit/s) sur l'enveloppe spectrale et le ratio tonales/bruit. Lors du décodage, la structure fine du spectre haute fréquence est synthétisée par réplication du spectre basse fréquence, puis l'enveloppe spectrale et les proportions relatives des tonales et du bruit sont corrigées en exploitant cette information. Cette technique permet au HE-AAC d'atteindre une qualité transparente à 24 kbit/s en mono, contre 64 kbit/s pour AAC.

La réplication spectrale ne tient pas compte de la fréquence fondamentale des sons harmoniques, ce qui rompt l'harmonicité (voir figure 2.6) et peut générer des phénomènes de battement et de rugosité liés aux positions relatives des tonales [6, 23]. Pour des sons tonals non-harmoniques (ex. : cloches), la SBR place les tonales haute-fréquence à des fréquences complètement différentes de celles du signal original.

Une solution a été proposée par [17], consistant à remplacer la réplication spectrale par un étalement des basses vers les hautes fréquences utilisant des vocodeurs de phase. Cette technique permet de préserver l'harmonicité, ce qui améliore la qualité audio, mais une part importante des harmoniques ne sont pas restaurées, de sorte que le timbre est mal reproduit. D'autre part, elle peut générer des pré- et des post-échos pour des signaux harmoniques à caractère percussif (ex. guitare). Elle a néanmoins été intégrée dans la norme USAC en 2011.

**Comme pour la réduction de pré-écho, nous avons d'abord conçu un système de correction des tonales fondé sur la transmission, par un canal auxiliaire, d'une information minimale, puis nous avons étudié comment adapter ce système dans le cas où le canal auxiliaire est un tatouage audio.**

Le système proposé est représenté sur la figure 2.5. Lors du codage, les trames tonales sont détectées et, le cas échéant, les fréquences des tonales sont estimées à la fois dans le signal original et dans sa future version décodée. L'erreur commise sur chaque tonale est transmise *via* le canal auxiliaire. Après le décodage, dans les trames détectées comme tonales, les fréquences des tonales sont de nouveau estimées, puis

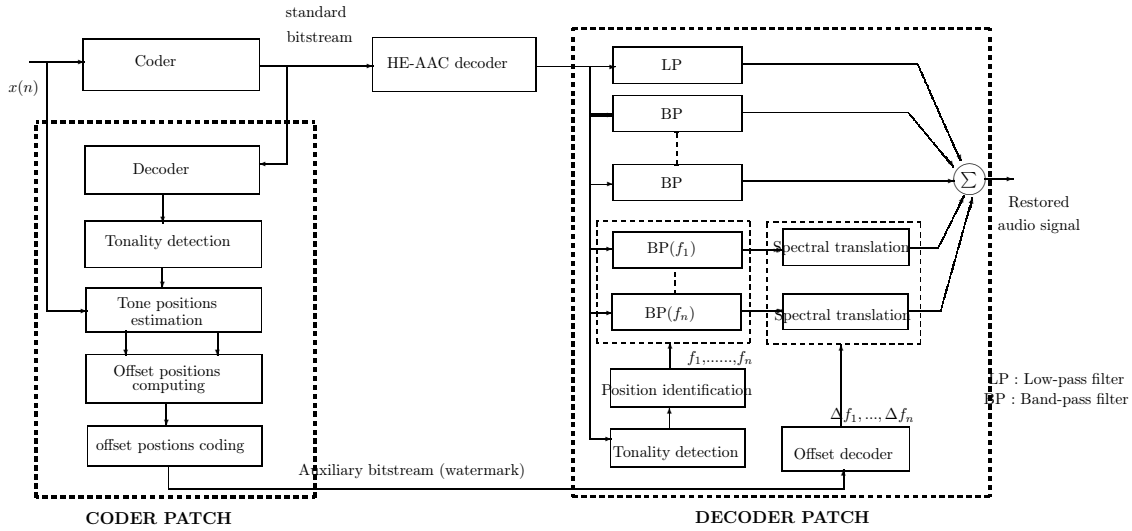



FIGURE 2.5 – Schéma général du système de correction de tonales.

corrigées grâce à l'information auxiliaire transmise. Ce système nécessite un décodage « à blanc » lors du codage, ce qui peut être vu comme une source de complexité et de retard. C'est néanmoins ce qui permet de transmettre non pas les positions des tonales, mais l'erreur qui sera commise par le SBR, solution plus économique en termes de débit d'information à transmettre. Nous détaillons ci-après les différents composants du système.

**Détection des tonales -** Une trame est caractérisée comme tonale si son *indice de tonalité* [11] dépasse un certain seuil. Cet indice repose la mesure de platitude spectrale, c'est-à-dire le rapport entre les moyennes géométrique et arithmétique du spectre, ce rapport étant proche de 1 pour le bruit et proche de 0 pour les signaux tonals (spectre parcimonieux).

**Estimation des fréquences des composantes tonales -** Nous avons proposé une méthode inspirée du modèle psychoacoustique de la norme MPEG-1, fondée sur la comparaison de l'amplitude de chaque composante avec celles de ses voisines. La simple application de ce modèle conduisait à la non-détection de tonales trop proches dans le signal décodé (résolution fréquentielle insuffisante) et à des fausses détections. Notre méthode repose sur une augmentation de la taille de la fenêtre d'analyse (amélioration de la résolution fréquentielle), un lissage du spectre par un filtre médian (élimination des pics parasites) et un seuillage du spectre par l'enveloppe spectrale calculée selon un modèle auto-régressif d'ordre 15. La figure 2.6 illustre l'efficacité de cette méthode à la fois sur le signal original et le signal codé-décodé. En appliquant la méthode de base MPEG-1, les paires de pics entourées auraient été identifiées comme un seul pic et plusieurs des maxima sous l'enveloppe spectrale auraient été identifiés à tort comme des tonales.

**Codage des décalages fréquentiels -** Chaque tonale du signal codé-décodé est appariée à une tonale du signal original et l'erreur de fréquence est codée selon un codage de Gray, avec un nombre de bits déterminé par le débit du canal auxiliaire et le nombre maximal de tonales à corriger. La dynamique de valeurs à considérer

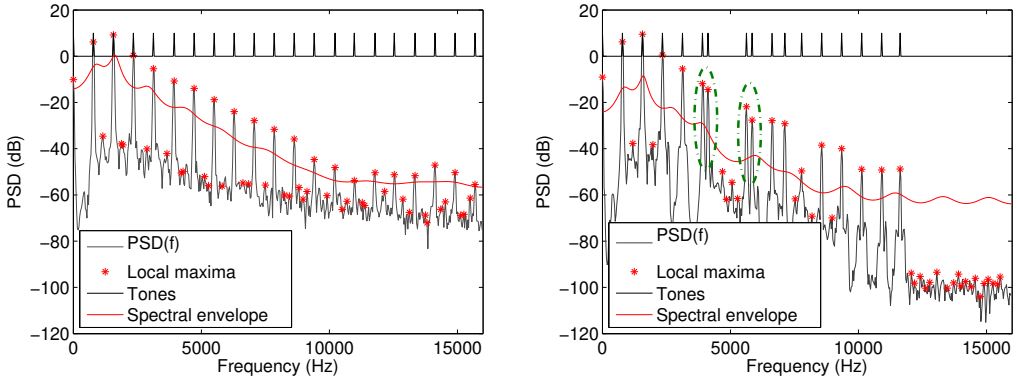



FIGURE 2.6 – Densité spectrale de puissance d'une séquence de 1024 ech. d'un signal de trompette échantillonné à 32 kHz, identification des maxima locaux et détection des composantes tonales par seuillage par l'enveloppe spectrale. (a) signal original - (b) signal codé/décodé par le HE-AAC à 16 kbit/s.

pour la quantification est l'intervalle $[-f_0; f_0]$, où $f_0$ est la fréquence fondamentale pour les signaux harmoniques et l'écart fréquentiel maximal entre deux tonales dans les basses-fréquences pour les signaux tonals non-harmoniques.

**Translations spectrales -** Le signal décodé passe à travers un banc de filtres divisant le signal d'une part en bandes de 100 Hz centrées autour des fréquences des tonales générées par la SBR, d'autre part en bandes contenant le reste du spectre. Chacune des bandes de la première catégorie subit une modulation à bande latérale unique (BLU) qui permet de translater la fréquence de la tonale selon l'erreur indiquée par l'information auxiliaire.

Le système proposé restaure bien les tonales à leurs positions originelles. Nous avons évalué l'effet perceptif par des mesures de rugosité [28] : pour les différents instruments testés, notre méthode rapproche la rugosité de celle du son original.

Dans un second temps, nous avons considéré l'utilisation d'un tatouage audio comme canal auxiliaire, selon la même démarche que dans la section 2.1. Pour cela, nous avons tracé d'une part les courbes de rugosité en fonction du TEB du canal auxiliaire pour plusieurs instruments, d'autre part les courbes de TEB de la détection du tatouage en fonction du débit d'insertion, pour un codec AAC+ à 16 et 20 kbit/s, avec le tatouage par étalement de spectre présenté en 1.4. La rugosité reste au même niveau tant que le TEB est inférieur à $10^{-2}$, ce qui nécessite de ne pas dépasser un débit de tatouage de 50 bit/s. Or la correction présentée *supra* suppose un débit d'information auxiliaire de l'ordre du kbit/s. Nous avons donc adapté la correction à la contrainte de débit en regroupant les trames traitées par notes, les changements de note étant détectés *via* les variations d'énergie par l'algorithme du codeur AAC [1] utilisé dans la section 2.1 pour la détection des attaques. Cette solution suppose (i) que les changements de note soient caractérisés par des attaques ; (ii) que les notes aient une durée suffisamment longue pour respecter le débit de tatouage ; (iii) que les erreurs de position des tonales provoquées par le codage-décodage soient les mêmes sur toutes les trames d'une note. Toutes ces conditions ne sont pas nécessairement

vérifiées. Sur le corpus testé, les mesures de rugosité sont cependant très proches des précédentes.

**Le tatouage réflexif permet donc, sous certaines conditions, de restaurer les fréquences originelles des tonales modifiées par une compression audio bas-débit à réplication de bande spectrale et, ainsi, de corriger la rugosité du signal codé-décodé.** Un exemple de correction est présenté sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/HDR/`.

De manière indépendante et parallèle, les auteurs de [17] ont également travaillé sur une amélioration du codec AAC+ fondée sur une translation spectrale par modulation BLU, la *continuous modulated bandwidth extension, CM-BWE* [18]. Il s'agit là non pas d'une correction externe au codec, mais d'un remplacement de la SBR consistant à synthétiser les hautes fréquences par modulation BLU des basses fréquences, avec une fréquence de modulation choisie de manière à préserver l'harmonicité. Cette méthode, qui a l'avantage de ne pas nécessiter d'information auxiliaire, a fait perdre de l'intérêt à notre proposition. Elle ne résout toutefois pas le problème des signaux tonals non-harmoniques.

## 2.3 Séparation de sources informée (par tatouage)

**Collaborations :**

— João-Marcos Travassos Romano, DSPcom, Unicamp

— Everton Nadalin, DSPcom, Unicamp

— Ricardo Suyama, CECS, UFABC

**Projet :** Compest (2012-2016)

Une première version du projet WaRRIS avait été conçue en 2005 avec des chercheurs du GIPSA-Lab (Grenoble) et du LaBRI (Bordeaux) avec un objet plus large [4], incluant notamment ce qui donnera lieu au projet DReaM, [5] à savoir *l'écoute active* : il s'agit de permettre à l'auditeur d'un mélange stéréo de manipuler indépendamment chaque instrument, ce qui nécessite une étape de séparation de sources. Le mélange étant sous-déterminé et convolutif, les performances des algorithmes de séparation de sources (voir [29]) sont insuffisantes pour l'usage visé. Le projet DReaM propose donc de **tatouer le mélange stéréo par des informations facilitant la séparation, d'où le nom de *séparation de sources informée* (*informed source separation*, ISS) [13].** Comme dans le projet WaRRIS, les sources originales sont accessibles, si chaque instrument a été enregistré séparément en studio avant le mixage. Le processus est schématisé par la figure 2.7.

Plusieurs techniques ont été proposées vers 2010. L'une [21] est fondée sur le principe de parcimonie conjointe de la représentation temps-fréquence des signaux audio : dans chaque région du plan temps-fréquence, rarement plus de deux sources sont actives simultanément, de sorte qu'il suffit de transmettre la matrice d'activité de chaque source et la matrice de mélange (supposé instantané) pour que le

4. Soumis à l'ANR dans le cadre de l'appel à projets jeunes chercheurs, il n'avait pas été retenu car sur-dimensionné pour cet appel, d'où la scission en deux projets distincts en 2006, WaRRIS et DReaM.

5. `https://dream.labri.fr/`

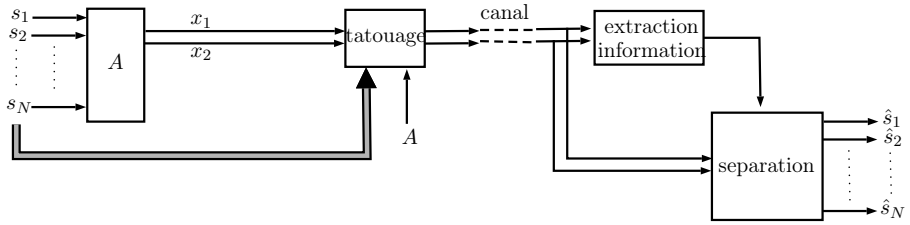



FIGURE 2.7 – Schéma de principe de la séparation de source informée sur un mélange stéréo $(x_1, x_2)$ de $N$ sources $s_1 \dots s_N$, avec une matrice de mélange $A$.

problème de séparation de sources se simplifie en la séparation d'un mélange stéréo de deux sources connaissant la matrice de mélange. Sur la base d'une modélisation des sources comme des processus gaussiens localement stationnaires, une séparation fondée sur un filtrage de Wiener est proposée dans [14], utilisant comme information auxiliaire les spectrogrammes de puissance des sources et les réponses fréquentielles des différents canaux source-capteur. De manière générale, les méthodes proposées (voir aussi [5, 27]) supposent la transmission d'une ou plusieurs de ces informations : spectrogrammes des sources, matrice temps-fréquence d'activité, filtres de mélange. Ainsi, le débit de tatouage requis est de l'ordre de la dizaine [21], voire de la centaine [14] de kbit/s. De fait, ces méthodes n'envisagent le tatouage que sur des signaux transmis dans des formats de compression sans perte, les débits nécessaires n'étant pas accessibles à des techniques robustes au codage perceptif.

En 2012, j'ai proposé à l'équipe DSPcom de l'Université de Campinas d'associer ses compétences en séparation de sources à l'expérience du projet WaRRIS pour **proposer une séparation de sources informée à faible débit de tatouage, robuste à la compression audio.**

Nous sommes partis de la méthode la moins gourmande en débit de tatouage, celle de Parvaix et Girin [21], améliorée par Pinel [22] : chaque composante temps-fréquence d'une source sous le seuil de masquage de celle-ci est supprimée, ce qui permet de mieux vérifier l'hypothèse de [21] selon laquelle au maximum deux sources sont actives simultanément en un instant et une fréquence. Reste à transmettre les matrices de dominance, qui indiquent dans le même plan temps-fréquence que le spectrogramme quelles sont les deux sources dominantes. Il peut s'agir soit d'une matrice binaire par source, indiquant pour chaque couple temps-fréquence$(t, f)$ si cette source est l'une des deux dominantes, soit d'une matrice unique indiquant pour chaque $(t, f)$ quelles sont les deux sources dominantes.

Pour adapter le débit de cette transmission à celui permis par un tatouage, une première idée repose sur la compression de ces matrices comme illustré par la figure 2.8 : un bloc vertical est comprimé *via* une méthode efficace (par exemple JBIG pour les matrices binaires) et le message binaire résultant est codé en contraignant un bloc adjacent de la matrice de dominance selon un alphabet de motifs [6]. Ce codage par contrainte est permis par ce que d'une part plusieurs motifs sont associés à chaque symbole, d'autre part de nombreux coefficients de cette matrice, indiquant la présence d'aucune ou d'une seule source, peuvent être remplacés par des coefficients indiquant la présence de deux sources. Le bloc ainsi modifié est codé dans un

6. D'où le terme d'auto-encodage, proposé avant qu'il ne devienne à la mode avec une autre signification dans le contexte des réseaux de neurones.

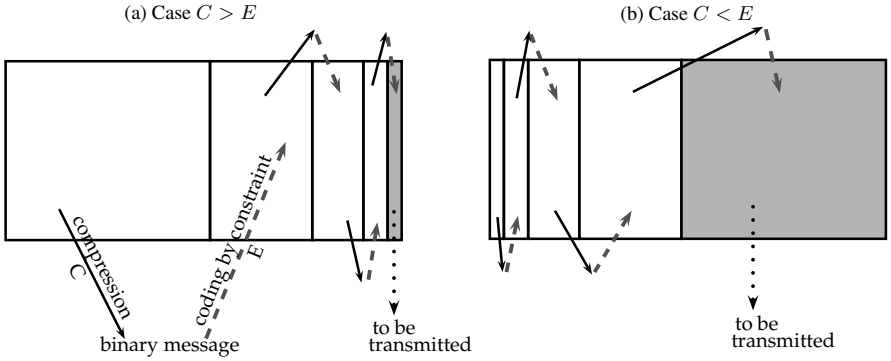



FIGURE 2.8 – Auto-encodage par contrainte d'une matrice de dominance, avec un taux de compression $C$ et un taux d'expansion $E$.

troisième bloc de la même manière, et ainsi de suite. Finalement, seul le dernier bloc d'information nécessite d'être transmis.

Pour que le gain en terme de débit de tatouage soit intéressant, il est nécessaire que le rapport entre le taux de compression $C$ et le taux d'expansion $E$ soit inférieur à 1 ou très légèrement supérieur. Cependant, le codage par contrainte d'un bloc augmente son entropie, ce qui réduit le taux de compression possible. D'autre part, alors que JBIG peut atteindre des taux de compression de l'ordre de 20 pour des images noir et blanc parcimonieuses et structurées (scan d'un manuscrit par exemple), ce taux descend à 2 ou 3 pour les matrices binaires de dominance. Cette piste a donc été écartée.

Nous avons alors suivi le principe consistant à ne transmettre que ce qui ne peut pas être estimé en réception, le rôle de l'information auxiliaire étant de corriger les erreurs d'estimation. Pour cela, nous considérons une séparation par analyse en composantes parcimonieuses (*Sparse Component Analysis*, SCA) [2]. Nous avons préféré la SCA locale à la SCA globale : celle-ci est en effet plus complexe, pour un gain en performance dont l'intérêt est limité ici, puisque les erreurs d'estimation seront compensées par l'information auxiliaire. Pour chaque coefficient temps-fréquence du mélange stéréo, la SCA locale estime l'ensemble de 0 à 2 sources dominantes et, dans le cas de deux sources, sépare celles-ci en appliquant l'inverse de la matrice de mélange (supposée connue) restreinte à ces sources. Ainsi, comme dans la section 2.2, la séparation est faite « à blanc » au niveau de l'émetteur, pour déterminer les erreurs d'estimation de sources dominantes qui seront commises en réception. **L'information auxiliaire transmise n'est donc plus la matrice de dominance binaire de chaque source, mais la matrice d'erreur correspondante. Le processus complet est résumé par la figure 2.9.**

Alors que les matrices de dominance binaire, dans un mélange de 3 sources, comportent de 20 à 40 % de 1, cette proportion est réduite à 10 % dans les matrices d'erreur. Cette parcimonie reste cependant insuffisante : pour une fréquence d'échantillonnage de 44,1 kHz, l'information binaire de dominance est de 44,1 kbit/s ; un codage entropique de l'erreur permet d'envisager une transmission d'environ 6 kbit/s par source, ce qui reste largement au-dessus des capacités d'un tatouage robuste à la compression audio.

Une autre difficulté est que la matrice de dominance ne peut être estimée que sur le mélange avant compression audio, puisque celle-ci est une opération non linéaire.

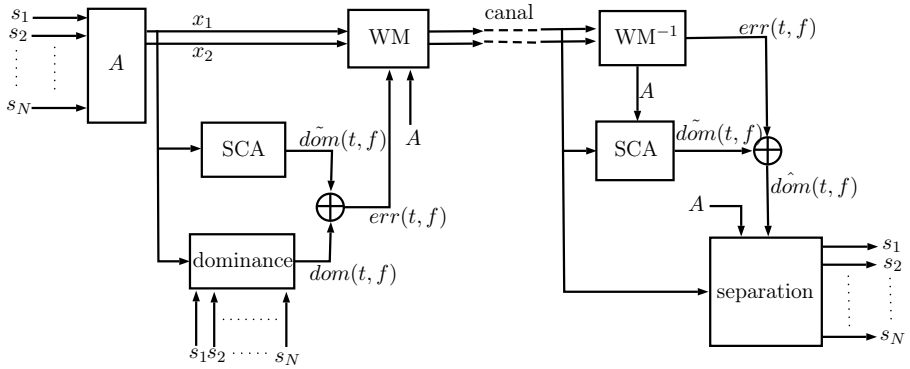



FIGURE 2.9 – Pour un mélange stéréo $(x_1, x_2)$ de $N$ sources $s_1 \dots s_N$, séparation de sources informée par transmission (par tatouage, noté WM) de la matrice de mélange $A$ et de l'erreur d'estimation des sources dominantes dans le plan temps-fréquence $err(t, f)$.

Ainsi, lorsque la chaîne de traitement inclus un codage-décodage audio, la qualité des signaux séparés n'est pas meilleure avec la méthode proposée (ou en utilisant directement les matrices de dominance) qu'avec une SCA sans information auxiliaire, ce qui indique que les matrices de dominance dans le mélange codé-décodé ne sont pas celles du mélange avant codage.

Face à ces obstacles, nous nous sommes tourné vers une nouvelle approche, celle du tatouage dopant, que nous avions déjà expérimentée avec succès pour la séparation de sources et pour d'autres applications : cette méthode sera présentée dans le chapitre suivant.

## 2.4 Conclusion

Nous avons montré qu'il est possible de tatouer un signal audio par des informations permettant sa propre restauration en sortie d'un canal de communication dégradé. L'application de ce principe est tributaire de l'état de l'art en tatouage audio : tandis que le compromis débit-inaudibilité-robustesse de l'état de l'art s'est avéré suffisant pour nos deux applications liées au codage audio à bas-débit (restauration des attaques et correction des fréquences des composantes tonales haute-fréquence), d'autres applications, comme la séparation de sources informée, peuvent nécessiter des débits de tatouage très supérieurs aux débits usuels assurant une robustesse suffisante à une dégradation légère telle que la compression audio.

Les applications étudiées, en tant que « patchs » correctifs de codecs audio, se heurtent sur le long terme à une limite inhérente à ce type de solution : l'évolution des codecs peut les rendre obsolètes. On l'a vu pour la correction des tonales : le fait que le défaut corrigé soit très circonscrit facilite certes une correction *via* une information auxiliaire suffisamment légère pour être transmise par tatouage, mais aussi une correction de la conception du codec lui-même. Pour la séparation de sources informée, parallèlement aux travaux de la communauté séparation de sources, la communauté codage a développé le codage spatial d'objets audio (*Spatial Audio Object Coding*, SAOC [8]), consistant à ajouter au codage d'un mélange audio une information auxiliaire permettant la séparation. Dès lors que la séparation de source informée est prévue dans le codage audio lui-même, elle pert de son intérêt en tant que sujet de recherche autonome. Notons que les deux approches ont été unifiées sous

le nom de séparation de sources informée fondée sur le codage (*coding-based informed source separation*, CISS [20]), qui sort du paradigme du tatouage réflexif.

L'intérêt des travaux présentés dans ce chapitre, au-delà des résultats, réside aussi dans le chemin suivi, qui a permis d'explorer de nombreux outils du traitement des signaux audio, utiles pour la suite. À cet égard, la thèse de Imen Samaali, tout en se concentrant sur des détails du codage audio, est le contraire d'une hyperspécialisation, par la variété des méthodes qu'elle a manipulées : codage, modèles psycho-acoustiques, tatouage, méthodes algébriques, représentations des signaux... Cet encadrement s'inscrit bien dans une conception du doctorat comme période de formation, au-delà de la seule production au bénéfice des encadrants.

Par la suite, nous étendons le principe du tatouage réflexif. Si l'on revient à la définition du tatouage, tatouer le signal pour en faciliter le traitement revient à modifier imperceptiblement le signal pour en faciliter le traitement, sans qu'il soit nécessaire que cette modification consiste en l'insertion d'une information explicite : ce sera l'approche étudiée dans les chapitres III et IV.

## 2.5 Références bibliographiques


[1] 3GPP. *Specification series, General audio codec audio processing functions ; Enhanced aacPlus general audio codec ; Encoder specification ; Advanced Audio Coding (AAC) part*, 2004.

[2] P. Comon and P. Jutten. *Handbook of Blind Source Separation.* Academic Press, 2010.

[3] M. Fliess, C. Join, and M. Mboup. Algebraic change-point detection. *Applicable Algebra in Engineering, Communication and Computing*, 21(2) :131–143, 2010.

[4] B. Geiser, F. Mertz, and P. Vary. Steganographic packet loss concealment forwireless voip. In *ITG Conference on Voice Communication [8. ITG-Fachtagung]*, pages 1–4, 2008.

[5] S. Gorlow and S. Marchand. Informed Source Separation : Underdetermined Source Signal Recovery from an Instantaneous Stereo Mixture. In *Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2011)*, pages 309–312, New Paltz, États-Unis, Oct. 2011.

[6] H. v. Helmholtz. On the sensations of tone (AJ Ellis, trans.). *Braunschweig : Vieweg & Son.(Original work published 1863)*, 1954.

[7] J. Herre. Temporal noise shaping, quantization and coding methods in perceptual audio coding : A tutorial introduction. *AES 17th Conference High Quality Audio Coding*, 1999.

[8] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh. MPEG spatial audio object coding—the ISO/MPEG standard for efficient coding of interactive audio scenes. *J. Audio Eng. Soc*, 60(9) :655–673, 2012.

[9] R. Huber and B. Kollmeier. PEMO-Q–a new method for objective audio quality assessment using a model of auditory perception. *IEEE trans. on audio, speech and language processing*, 14(6) :1902 – 1911, November 2006.

[10] ITU. *ITU-R Rec. BS. 1387 : Method for objective measurement of perceived audio quality*, 1998.

[11] J. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE J. on selected areas in communications*, 6(2) :314–323, Feb. 1988.

[12] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1) :84–95, 1980.

[13] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard. Informed source separation : a comparative study. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[14] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8) :1937 – 1949, 2012.

[15] D. P. M. Athineos. Autoregressive modeling of temporal envelopes. *IEEE transaction on signal processing*, 55(11), 2007.

[16] M. Mboup, C. Join, and M. Fliess. A delay estimation approach to change-point detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[17] F. Nagel and S. Disch. A harmonic bandwidth extension method for audio codecs. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 145–148, 2009.

[18] F. Nagel, S. Disch, and S. Wilde. A continuous modulated single sideband bandwidth extension. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 357–360, 2010.

[19] P. Noll. *MPEG Digital Audio Coding Standards*. CRC Press LLC, 2000.

[20] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation : Nonnegative tensor factorization approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8) :1699–1712, 2013.

[21] M. Parvaix, L. Girin, and J. Brossier. Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding . *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6) :1721 – 1733, Aug. 2011.

[22] J. Pinel and L. Girin. "Sparsification" of audio signals using the MDCT/IntMDCT and a psychoacoustic model – application to informed audio source separation. In *Proc. of the 42nd Audio Engineering Society Conference : Semantic Audio*, Ilmenau, Germany, 2011.

[23] R. Plomp and W. J. M. Levelt. Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4) :548–560, 1965.

[24] M. J. S. Larbi. Audio watermarking : A way to stationarize audio signals. *IEEE Trans. Signal Processing*, 53(2) :816–823, 2005.

[25] A. Sagi and D. Malah. Bandwidth extension of telephone speech aided by data embedding. *EURASIP Journal on Advances in Signal Processing*, 2006.

[26] S. Sarreshtedari, M. A. Akhaee, and A. Abbasfar. A watermarking method for digital speech self-recovery. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11) :1917–1925, 2015.

[27] N. Sturmel and L. Daudet. Informed source separation using iterative reconstruction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1) :178–185, 2013.

[28] P. N. Vassilakis. Sra : A web-based research tool for spectral and roughness analysis of sound signals. In *Proceedings SMC'07, 4th Sound and Music Computing Conference*, page 319–325, 2007.

[29] E. Vincent and Y. Deville. *Handbook of Blind Source Separation*, chapter Audio applications, pages 779–820. Academic Press, 2010.

[30] E. Zwicker and R. Feldtkeller. *The ear as a communication receiver*. Acoustical Society of America, New York, 1999.

# Chapitre 3

# Du bruit pour doper le signal

## Le tatouage dopant

**Sommaire**



Dans ses travaux de thèse sur le tatouage audio [21], Sonia Larbi a remarqué qu'un tatouage par étalement de spectre rend le signal hôte plus stationnaire. Elle a montré avec Mériem Jaïdane [37] que ce tatouage améliorait les performances en régime stationnaire d'un annuleur d'écho piloté par le signal hôte : l'ERLE (*Echo Return Loss Enhancement*) est amélioré d'environ 5 dB. Cette amélioration s'explique aisément par ce que les algorithmes adaptatifs utilisés en annulation d'écho sont sensibles à la stationnarité du signal [23]. Auparavant, Valérie Turbin et André Gilloire avaient montré [10] que l'insertion d'un bruit inaudible dans le signal de parole permet d'améliorer les performances d'un annuleur d'écho stéréophonique en réduisant la corrélation entre les deux voies.

Revenons à l'objectif exprimé en introduction d'un tatouage facilitant un traitement du signal hôte : ici, le tatouage est intéressant non seulement par l'information

binaire qu'il véhicule, mais aussi en ce qu'il transforme les propriétés du signal et ainsi facilite l'annulation d'écho. Nous avons alors exploré l'idée de *tatouage dopant* : il s'agit de concevoir **un tatouage qui modifie les propriétés du signal de manière à faciliter un traitement ultérieur. Le tatouage, dans ce cas, ne véhicule plus nécessairement d'information explicite** (message binaire), mais consiste en un simple bruit additif comme dans [10] (dans l'esprit du tatouage additif) ou en toute autre modification inaudible du signal audio (dans l'esprit du tatouage substitutif).

Cette proposition est motivée de manière plus générale par ce que de nombreux algorithmes de traitement du signal, que nous préciserons dans la suite du chapitre, reposent sur des hypothèses qui sont rarement vérifiées pour les signaux audio - stationnarité, blancheur, gaussianité - ou insuffisamment vérifiées - non-gaussianité, parcimonie, indépendance. Plus les signaux réels s'éloignent de ces propriétés « idéales », plus les performances des algorithmes se dégradent. Usuellement, la recherche consiste à adapter les algorithmes aux propriétés réelles des signaux, au prix d'une complexité accrue. L'approche développée ici consiste au contraire à rendre les algorithmes de base performants en adaptant le signal aux hypothèses requises, au prix d'une distorsion qui doit rester inaudible.

Outre les travaux de Sonia Larbi, cette idée de dopage des signaux a aussi été inspirée par le phénomène de résonance stochastique [4], où une augmentation du niveau de bruit peut, dans certains contextes (bruit non gaussien et mélange non-linéaire, par exemple), améliorer les performances d'un estimateur ou d'un détecteur optimal.

Le chapitre est organisé comme suit. La première section est consacrée à l'égalisation d'histogramme dans différents contextes où le signal original est disponible et où les algorithmes de traitement requièrent une distribution particulière des échantillons temporels ou temps-fréquence : identification de systèmes non-linéaires ; séparation de sources informée ; application du théorème de quantification. Nous nous intéressons aussi au lien entre histogramme et séparation de sources dans la section 2, mais en considérant ici la parcimonie jointe des sources, c'est-à-dire le faible recouvrement entre les sources dans le domaine temps-fréquence. Enfin, nous montrons dans la section 3 comment il est possible de débruiter un signal en l'ayant, avant transmission dans un canal bruité, approché par une suite de combinaisons linéaires de codes correcteurs d'erreurs.

## 3.1 Égalisation d'histogramme

**Collaborations :**

- Mériem Jaïdane, Sonia Larbi et Monia Turki (U2S, ENIT)
- Hmaied Shaiek (Cabasse, Brest)
- João-Marcos Romano et Everton Nadalin (DSPcom, Unicamp)
- Ricardo Suyama (CECS, Universidade Federal do ABC)

**Encadrements :**

- thèse de Imen Mezghani, en co-tutelle Paris 5 - ENIT (2005-2010), co-encadrée avec Mériem Jaïdane, Sonia Larbi et Monia Turki (U2S, ENIT) ;

— stage de M2 (Université Pierre et Marie Curie) de Houssem Halalchi, 2008, co-encadré avec Mériem Jaïdane

— stage de M2 (Université Pierre et Marie Curie) de Abdelmoumène Mékentichi, 2009, co-encadré avec Mériem Jaïdane

**Projets :** WaRRIS, CMCU2004 et Compest (2005-)

**Publications :**

[HMJ09] H. Halalchi, G. Mahé, and M. Jaïdane. Revisiting quantization theorem through audiowatermarking. In *Proc. ICASSP 2009*, pages 3361–3364, Taipei, Taïwan, 2009.

[MJ18] Gaël Mahé and Mériem Jaidane. Perceptually Controlled Reshaping of Sound Histograms. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(9) :1671 – 1683, September 2018.

[MMJS+07] I. Marrakchi, G. Mahé, M. Jaidane-Saidane, S. Djaziri-Larbi, and M. Turki-Hadj Alouane. Gaussianisation method for identification of memoryless nonlinear audio systems. In *Proceedings of the 15th European Signal Processing Conference (Eusipco 2007)*, pages 2316–2320, 2007.

[MMMDL+14] I. Mezghani-Marrakchi, G. Mahé, S. Djaziri-Larbi, M. Jaidane, and M. Turki-Hadj Alouane. Nonlinear audio systems identification through audio input gaussianization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1) :41–53, january 2014.

[MNR12] G. Mahé, E.Z. Nadalin, and J.M.T. Romano. Doping audio signals for source separation. In *Proceedings of the 20th European Signal Processing Conference (Eusipco 2012)*, pages 2402–2406, Bucarest, Romania, August 2012.

[MNSR14] Gaël Mahé, Everton Nadalìn, Ricardo Suyama, and João Romano. Perceptually controlled doping for audio source separation. *EURASIP Journal on Advances in Signal Processing*, 2014(1) :27, march 2014.

### 3.1.1 Des contextes où la distribution du signal n'est pas satisfaisante

#### Identification de systèmes audio non-linéaires

Les microphones, amplificateurs et haut-parleurs souffrent de fonctionnements non-linéaires, d'origine électrique, mécanique ou acoustique. De ce fait, ils peuvent être modélisés soit par des systèmes polynômiaux (pour le cas sans mémoire), soit par des filtres de Volterra (pour le cas avec mémoire, notamment amplificateurs et haut-parleurs [16, 18, 42]). L'identification de ces filtres ou la caractérisation des systèmes (par leur réponse fréquentielle et leur taux de distorsion harmonique) utilise classiquement des signaux synthétiques mathématiquement adaptés : bruit blanc gaussien [39, 29], multitons [19], sinusoïdes glissantes [30], séquences à longueur maximale [17]. Pourtant, comme l'a montré Klippel [20], le comportement d'un haut-parleur dépend du signal d'entrée, de sorte que l'identification d'un système audio non-linéaire devrait être réalisée avec les signaux d'usage, parole ou musique. Mais

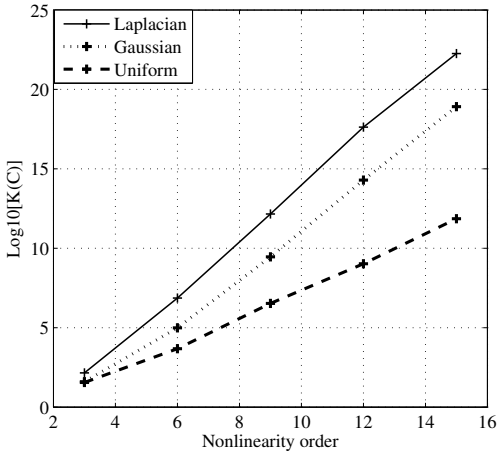



FIGURE 3.1 – Pour un système polynomial, conditionnement $K(\mathbf{C_x})$ en fonction de l'ordre de non-linérarité, pour différentes densités de probabilité.

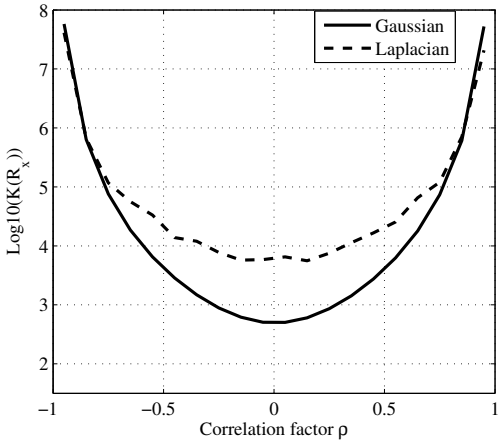



FIGURE 3.2 – Pour un système de Volterra de mémoire $M = 2$ et d'ordre de non-linéarité $N = 4$, conditionnement $K(\mathbf{C_x})$ selon le facteur de corrélation $\rho$ pour deux lois de probabilité, pour un signal AR1.

ceux-ci ont toutes les « mauvaises » propriétés pour les algorithmes classiques d'identification de système : non-gaussianité, non-stationnarité, forte corrélation.

Dans le cadre de la thèse de Imen Mezghani, nous avons étudié le lien entre la densité de probabilité des signaux et les performances d'une identification optimale [MMJS+07, MMMDL+14]. Dans le cas d'un modèle polynômial d'ordre $N$, la relation entrée-sortie est de la forme :

$$y_k = A^\top X_k, \quad \text{avec } X_k = (x_k, x_k^2, \dots x_k^N)^\top \tag{3.1}$$

Les performances d'identification dépendent du conditionnement de la matrice $C_x = \mathrm{E}[XX^\top]$. Comme indiqué par la figure 3.1, le conditionnement est d'autant meilleur que la distribution est plate et les écarts entre distributions augmentent avec l'ordre de non-linéarité. Pour une modélisation par un filtre de Volterra,

$$\begin{aligned} y_k &= A^T X_k, \\ &\text{avec } X_k \text{ constitué de tous les } x_k^{m_1} x_{k-1}^{m_2} ... x_{k-M+1}^{m_M} \\ &\text{tel que } m_1 + m_2 + ... + m_M \leq N \end{aligned} \tag{3.2}$$

Les performances d'identification dépendent du conditionnement de la matrice $R_x = \mathrm{E}[XX^\top]$. Comme démontré par [31], pour un processus indépendant et identiquement distribué, le conditionnement de $R_x$ augmente exponentiellement avec l'ordre de non-linéarité $N$ et la longueur $M$ de la mémoire, et a pour borne supérieure celui de $C_x$ à la puissance $N$. Pour un signal corrélé, ces propriétés sont difficiles à étudier théoriquement, nous avons donc montré expérimentalement l'influence de la distribution et de la corrélation sur le conditionnement de $R_x$, pour un signal auto-régressif d'ordre 1. La figure 3.2 confirme le résultat précédent pour les faibles corrélations, tandis que pour les fortes corrélations, la distribution importe peu ; dans tous les cas, la corrélation dégrade le conditionnement de $R_x$.

Une méthode d'identification a été proposée par [24] pour des signaux stationnaires, gaussiens et corrélés, passant par des étapes de prédiction et d'orthogonalisation avant l'identification elle-même. La gaussianité est nécessaire ici pour simplifier

l'orthogonalisation en utilisant des polynômes d'Hermite à la place de la procédure de Gram-Schmidt.

Dans tous les cas, la gaussianité du signal est donc souhaitable.

### Séparation de sources

Lorsque le nombre de capteurs est égal au nombre de sources, les méthodes de séparation fondées sur l'analyse en composantes indépendantes (ICA) supposent qu'au moins une des sources n'est pas gaussienne (théorème de Darmois). Sous cette hypothèse, un mélange déterminé instantané peut théoriquement être parfaitement séparé. En pratique, les performances de l'ICA se dégradent dès que la distribution des sources s'approche de la gaussianité. La figure 3.13 (points « SIR/SDR original » et « SAR original ») montre des performances très faibles pour des signaux à distribution gaussienne généralisée ayant un facteur de forme entre 1,4 et 2, dans le scénario pourtant le plus simple, à savoir la séparation d'un mélange stéréo de deux sources. La super-gaussianité des signaux audio peut donc être insuffisante pour l'ICA.

Lorsque le nombre de capteurs est inférieur au nombre de sources, la séparation utilise une analyse en composantes parcimonieuses (*Sparse Component Analysis*, SCA), fondée sur l'idée que les sources ont une distribution parcimonieuse dans un certain domaine de représentation. Pour les signaux audio, on exploite la parcimonie dans le domaine temps-fréquence. Cette propriété permet de considérer que, dans une région de l'espace temps-fréquence, le nombre de sources est inférieur au nombre de capteurs (parcimonie jointe). Cette hypothèse peut être exploitée pour identifier des sous-espaces contenant les sources et en déduire les paramètres du mélange [40, 41, 27] ou pour appliquer localement (dans l'espace temps-fréquence) des méthodes d'analyse en composantes indépendantes (ICA, adaptée aux mélanges avec autant de capteurs que de sources) [28, 26]. À défaut d'une parcimonie stricte au sens $\ell_0$, la SCA peut être facilitée par une parcimonie au sens $\ell_{\alpha\leq 1}$ (par exemple une distribution super-gaussienne).

Les performances de l'état de l'art de la SCA appliquée au signaux audio [44] restent cependant faibles, notamment parce que l'hypothèse de parcimonie est insuffisamment vérifiée.

### Théorème de quantification

Le théorème de quantification [47] est l'équivalent du théorème de Shannon pour les densités de probabilité. Il énonce que la densité de probabilité (ddp) d'un signal peut être reconstruite parfaitement à partir du signal quantifié si l'inverse du pas de quantification est supérieur à 2 fois la fréquence maximale de la fonction caractéristique. Au cours du stage de M2 de Houssem Halalchi, nous avons adapté ce théorème à la reconstruction des densités de probabilité ou des histogrammes empiriques de signaux quantifiés après une augmentation du pas de quantification [HMJ09].

Considérons un signal quantifié $x$ et supposons, sans perte de généralité, qu'il prenne ses valeurs dans $\mathbb{N}$. Sous-quantifier $x$ d'un facteur $K$ signifie arrondir chaque valeur de l'intervalle discret $\lfloor nK - K/2 + 1, nK + K/2 \rfloor$ à $nK$ ($n$ entier). La sous-quantification revient à une version discrète de ce que Widrow a appelé *l'échantillonnage de surface*, illustré par la figure 3.3 : la ddp $f_x$ est convoluée par une fenêtre

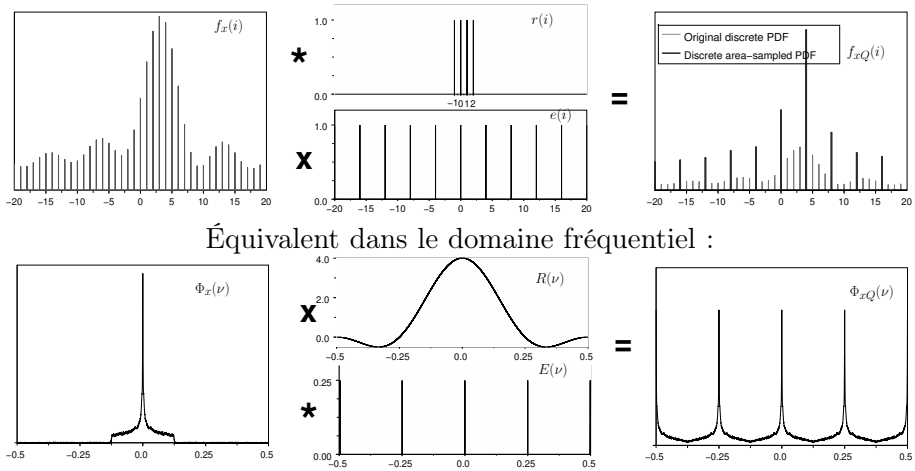



FIGURE 3.3 – Sous-échantillonnage de surface pour $K = 4$.

rectangulaire de largeur $K$ puis le résultat est multiplié par un train d'impulsions de période $K$, ce qui revient à exprimer la fonction caractéristique de $x_Q$ :

$$\Phi_{xQ}(\nu) = [\Phi_x(\nu)\, R(\nu)] * E(\nu) \tag{3.3}$$

avec :

$$R(\nu) = \begin{cases} K & \text{si } \nu \in \mathbb{Z} \\ \frac{sin(\pi K \nu)}{sin(\pi \nu)} \exp(j\pi\nu) & \nu \notin \mathbb{Z} \text{ et } K \text{ pair} \\ \frac{sin(\pi K \nu)}{sin(\pi \nu)} & \nu \notin \mathbb{Z} \text{ et } K \text{ impair} \end{cases} \tag{3.4}$$

$$E(\nu) = \frac{1}{K} \sum_{n=-\infty}^{\infty} \delta(\nu - \frac{n}{K}) \tag{3.5}$$

Dans le domaine fréquentiel, sous-quantifier d'un facteur $K$ revient à $K$-périodiser la fonction caractéristique multipliée par $R$. Nous avons ainsi formulé un **théorème de sous-quantification** :

**Théorème 3.1** *Si la fonction caractéristique $\Phi_x$ d'un signal quantifié $x$ est nulle pour $|\nu| > \frac{1}{2K}$ dans $[-\frac{1}{2}; \frac{1}{2}]$, $K \in \mathbb{N}$, alors la densité de probabilité de $x$ peut être restaurée à partir de celle du signal $x_Q$ résultant d'une sous-quantification de $x$ d'un facteur $K$. Dans le domaine fréquentiel, cette restauration s'exprime :*

$$\Phi_x(\nu) = \Phi_{xQ}(\nu) G(\nu) \tag{3.6}$$

*où $\Phi_{xQ}$ désigne la fonction caractéristique de $x_Q$ et $G(\nu) = 1/R(\nu)$ pour $|\nu| < \frac{1}{2K}$, 0 sinon.*

Dans le cadre du stage de M2 de Abdelmoumène Mékentichi, nous avons étendu cette étude à la reconstruction de la ddp jointe discrète $f_{x_1 x_2}$ de deux signaux quantifiés $x_1$ et $x_2$ après sous-quantification de facteurs respectifs $K_1$ et $K_2$.

Cette sous-quantification peut s'exprimer comme un *sous-échantillonnage de volume* de la ddp jointe discrète (version 2D du sous-échantillonnage de surface), consistant à convoluer $f_{x_1 x_2}$ par une fenêtre 2D rectangulaire puis multiplier le résultat

par un train bidimensionnel d'impulsions espacées de $(K_1, K_2)$. Dans le domaine fréquentiel, ces opérations sont équivalentes à :

$$\Phi^Q_{x_1x_2}(\nu_1,\nu_2) = [\Phi_{x_1x_2}(\nu_1,\nu_2)R_2(\nu_1,\nu_2)] * E_2(\nu_1,\nu_2) \tag{3.7}$$

avec :

$$R_2(\nu_1,\nu_2) = R(\nu_1)R(\nu_2) \tag{3.8}$$

$$E(\nu_1,\nu_2) = \frac{1}{K_1K_2}\sum_{n=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}\delta(\nu_1-\frac{n}{K_1})\delta(\nu_2-\frac{m}{K_2}) \tag{3.9}$$

On en déduit un **théorème de sous-quantification 2D** :

**Théorème 3.2** *Si la fonction caractéristique jointe $\Phi_{x_1x_2}$ de deux signaux quantifiés $x_1$ et $x_2$ est nulle pour $|\nu_i| > \frac{1}{2K_i}$ ($i \in \{1;2\}$, $K_i \in \mathbb{N}$) dans $[-\frac{1}{2};\frac{1}{2}]$, alors la densité de probabilité jointe de $(x_1, x_2)$ peut être restaurée à partir de celle des signaux $(x_1^Q, x_2^Q)$ résultant d'une sous-quantification de $x_1$ (resp. $x_2$) d'un facteur $K_1$ (resp. $K_2$). Dans le domaine fréquentiel, cette restauration s'exprime :*

$$\Phi_{x_1x_2}(\nu_1,\nu_2) = \Phi^Q_{x_1x_2}(\nu_1,\nu_2)G(\nu_1,\nu_2) \tag{3.10}$$

*où $G(\nu_1,\nu_2) = 1/R_2(\nu_1,\nu_2)$ pour $|\nu_1| < \frac{1}{2K_1}$ et $|\nu_2| < \frac{1}{2K_2}$, 0 sinon.*

Tout ce qui précède reste valable en considérant non pas la ddp mais l'histogramme empirique d'un signal de durée finie. Le théorème de sous-quantification permet de reconstruire l'histogramme originel d'un signal sous-quantifié et de déduire des propriétés statistiques d'un signal de sa version sous-quantifiée, mais pas de déquantifier le signal : la sous-quantification reste une perte d'information irréversible. Au-delà de la ddp jointe de deux signaux, il est cependant possible d'étendre ce théorème à la densité de probabilité jointe d'une séquence de $p$ échantillons successifs, ce qui ouvrirait la voie à une déquantification.

Toutefois, les conditions du théorème de sous-quantification sont rarement vérifiées, notamment pour des histogrammes empiriques, qui sont d'autant moins réguliers que la séquence analysée est courte, ce qui étale la fonction caractéristique vers les hautes fréquences.

### 3.1.2 Premières méthodes d'égalisation d'histogramme

Dans tout ce qui suit, nous noterons $x$ le signal original, $z$ le signal transformé (ou en cours de transformation). Pour un signal $s$, nous noterons $f_s$ son histogramme et $F_s$ son histogramme cumulé. Le spectrogramme d'un signal $s$ sera noté $S$, dont l'histogramme et l'histogramme cumulé du module seront notés respectivement $f_{|S|}$ et $F_{|S|}$. Nous noterons $f_{target}$ l'histogramme cible et $F_{target}$ l'histogramme cumulé cible.

Dans les cas présentés dans la sous-section 3.1.1, deux besoins apparaissent : soit modifier la forme générale de l'histogramme du signal, pour rendre celui-ci plus gaussien (identification de systèmes non-linéaires) ou plus parcimonieux (séparation de sources), soit modifier finement l'histogramme pour réduire le support de la fonction caractéristique (filtrage passe-bas en vue de l'application du théorème de sous-quantification). Pour chacun de ces objectifs nous avons proposé une méthode d'égalisation d'histogramme.

### Reformage global d'histogramme (HGR)

Le reformage global s'inspire de l'égalisation d'histogramme utilisée pour le rehaussement de contraste d'une image. Il s'agit de remplacer chaque échantillon $x(n)$ du signal par :

$$z(n) = F_{target}^{-1}\left(F_x\left(x(n)\right)\right) \tag{3.11}$$

La même équation s'applique pour des échantillons temps-fréquence.

L'application brute de cette transformation à chaque échantillon temporel [MMJS+07, MMMDL+14] se traduit par un bruit additif nettement audible. Aussi avons-nous limité la transformation aux segments voisés (pour la parole) ou tonals (pour la musique), moins sensibles, et limité l'amplitude maximale du bruit ajouté de telle sorte que sa variance soit inférieure ou égale à une variance cible, fixée empiriquement de manière à assurer l'inaudibilité du bruit de transformation. La gaussianisation de parole ou de musique selon cette méthode produit un signal certes pas tout à fait gaussien, mais dont le kurtosis se rapproche de 3.

L'application de la transformation 3.11 aux modules des échantillons temps-fréquence $|X(m, f)|$ [MNR12] revient à filtrer chaque $m^{\text{ème}}$ trame par un filtre de réponse fréquentielle $|Z(m, f)|/|X(m, f)|$, de sorte que la dégradation introduite n'est pas un bruit additif mais une distorsion spectrale variant dans le temps. Nous avons mis en œuvre cette transformation avec un histogramme cible gaussien généralisé de facteur de forme deux fois inférieur au facteur de forme du signal original, de manière à faciliter la séparation de sources par SCA. Les signaux ainsi transformés ont un facteur de forme nettement inférieur à celui de la distribution originelle, mais d'autant supérieur à la valeur visée que le signal est analysé et transformé sur une courte durée. Cette difficulté à atteindre le facteur de forme cible peut s'expliquer (i) par le recouvrement entre trames lors du filtrage, qui conduit à un spectrogramme légèrement différent de la cible ; (ii) par ce que la méthode d'égalisation d'histogramme utilisée est connue pour atteindre d'autant moins précisément la distribution cible que le nombre d'échantillons est faible. La dégradation des signaux, mesurée par PESQ [14] pour la parole, est presque toujours inaudible, mais la probabilité d'audibilité augmente en réduisant la durée du signal.

### Reformage local d'histogramme (HLR)

Le principe est de déplacer les échantillons entre des classes voisines de l'histogramme, des classes excédentaires (par rapport à l'histogramme cible) vers les classes déficitaires. Initialement, $z = x$. Puis, pour $j$ variant de la valeur minimale à la valeur maximale de $x$,

- Si $f_z(j) - f_{target}(j) = M > 0$, $M$ échantillons de $z$ de valeur $j$ sont sélectionnés aléatoirement. Chacun de ces échantillons prend la valeur $j + 1$, de sorte que $f_z(j) = f_{target}(j)$ et $f_z(j + 1) = f_z(j + 1) + M$.
- Si $f_z(j) - f_{target}(j) = -M < 0$, $M$ échantillons de $z$ de valeur $j + 1$ sont sélectionnés aléatoirement. Chacun de ces échantillons prend la valeur $j$. Si $f_z(j + 1) < M$, les échantillons manquants sont sélectionnés aléatoirement parmi ceux de valeur $j+2$, et ainsi de suite, jusqu'à atteindre $f_z(j) = f_{target}(j)$.

À la fin de l'algorithme, l'histogramme cible est exactement atteint.

Dans l'application visée, le filtrage passe-bas de l'histogramme, le bruit de transformation est d'autant plus fort que la fréquence de coupure est faible. La fréquence de coupure visée est $1/2K$, avec $K$ le facteur de sous-quantification prévu. Pour une durée de signal de 3 à 4 s, selon le type de signal et sa fréquence d'échantillonnage, une valeur maximale de $K$ entre 4 et 8 assure l'inaudibilité de la transformation.

La transposition de l'algorithme à un histogramme 2D (densité de probabilité jointe) se heurte à plusieurs difficultés :

— la quantification usuelle du signal original sur 16 bits implique des structures de données dont l'occupation mémoire dépasse les capacités courantes ;

— du fait des nombreuses valeurs nulles ou très faibles dans l'histogramme, le filtrage passe-bas crée un histogramme cible avec nombre important de valeurs inférieures à 1, voire négatives ;

— quand bien même l'histogramme filtré n'aurait que des valeurs entières positives, rien ne garantit qu'il puisse être représentatif d'une ddp jointe ;

— tout déplacement d'un couple d'échantillons $(x(n-1), x(n))$ dans l'histogramme agit simultanément sur $(x(n-2), x(n-1))$ et $(x(n), x(n+1))$.

Cette dernière difficulté a été résolue par l'algorithme séquentiel suivant. Initialement, $z = x$. Puis, pour chaque $n$, si $f_z(z(n), z(n-1))$ est excédentaire, on cherche le déplacement $\delta$ de $z(n)$ tel que

$$|z(n) + \delta - x(n)| \text{ minimal s.c.} \tag{3.12}$$

$$\begin{cases} f_z(z(n)+\delta, z(n-1)) \text{ déficitaire} \\ |f_z - f_{target}|(z(n+1), z(n)+\delta)) < |f_z - f_{target}|(z(n+1), z(n))) \end{cases} \tag{3.13}$$

Ce processus est répété jusqu'à la convergence de la distance de la variation totale entre $f_z$ et $f_{target}$. Contrairement à l'algorithme 1D, cet algorithme ne permet pas d'atteindre exactement l'histogramme cible, mais le support de la fonction caratéristique est bien réduit.

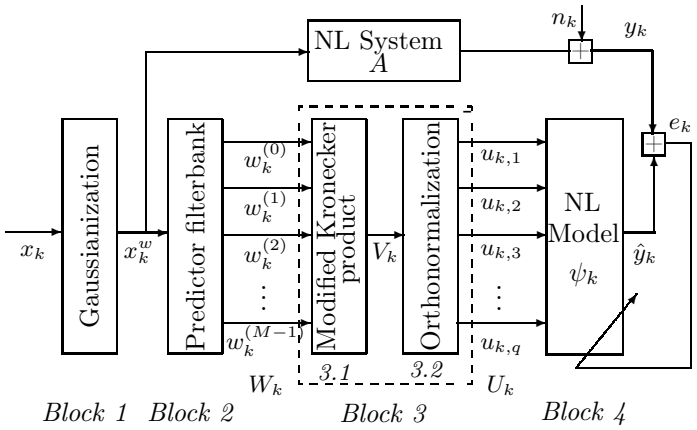
À partir de 2011, nos efforts se sont concentrés sur le contrôle perceptif des algorithmes de reformage des histogrammes 1D, de sorte que ces travaux sur les histogrammes 2D ont été mis de côté.

### 3.1.3 Utilisation des signaux à histogramme reformé

#### Identification de systèmes non-linéaires

La gaussianité donnée aux signaux audio par le reformage global d'histogramme nous a permis de proposer le schéma d'identification de système non-linéaire détaillé sur la figure 3.4 [MMMDL+14], inspiré de [24], qui comporte une étape de blanchiment suivie d'une orthonormalisation exploitant la gaussianité, avant l'identification adaptative elle-même *via* l'algorithme classique NLMS [1].

1. On pourrait questionner le choix d'une identification adaptative alors que la gaussianisation se fait sur des longues séquences, de l'ordre de la seconde, donc « hors-ligne » . Ce choix est motivé par la variabilité du système dans le temps, résultant à la fois de sa dépendance au signal et de son échauffement. La contrainte qui pèse sur la gaussianisation limite cependant les cas d'usage à l'identification d'un système diffusant des signaux audio préalablement enregistrés.




Le banc de filtres de prédiction calcule les erreurs de prédictions aux ordres 0 à $M-1$, avec $M$ la mémoire du système de Volterra. Cette étape ne concerne pas les systèmes sans mémoire et ne nécessite pas la gaussianité du signal.
Le but du bloc 3 d'orthonormalisation est de fournir à l'entrée du modèle non-linéaire (bloc 4) un vecteur orthonormal.
Pour les systèmes sans mémoire, on normalise $x_k^w$ par sa variance, estimée toutes les 10 à 30 ms. Comme $x_k$ est gaussien, le vecteur $X_k^w = (1, \tilde{x}_k^w, (\tilde{x}_k^w)^2, ..., (\tilde{x}_k^w)^N)^\top$, où $\tilde{x}_k^w$ désigne $x_k^w$ normalisé, peut être facilement orthonormalisé en utilisant les polynômes d'Hermite normalisés. On forme ainsi le vecteur $U_k = (\tilde{H}_0(\tilde{x}_k^w), \tilde{H}_1(\tilde{x}_k^w), ..., \tilde{H}_N(\tilde{x}_k^w))^\top$, avec $\tilde{H}_i$ le $i^{\text{ème}}$ polynôme d'Hermite normalisé, tel que $\mathrm{E}[U_k U_k^\top]$ est la matrice unité. On a la relation suivante

$$U_k = \mathbf{\Gamma_k} X_k^w, \tag{3.14}$$

où $\mathbf{\Gamma_k}$ est une matrice triangulaire inférieure de dimensions $(N+1)\times(N+1)$.
Pour les systèmes avec mémoire, le modèle de Volterra peut être remplacé par un modèle de Wiener qui, pour un signal gaussien, est une combinaison linéaire de produits de polynômes d'Hermite tels que le degré est inférieur ou égal à $N$, celui du modèle de Volterra. On forme alors un vecteur $U_k$ orthonormal composé de produits de facteurs de la forme $\widetilde{H}_j(\tilde{w}_k^{(i)})$, où $\tilde{w}_k^{(i)}$ est l'erreur de prédiction d'ordre $i$ normalisée par sa variance. Le vecteur $U_k$ peut s'écrire $U_k = \mathbf{Q}V_k$, avec $\mathbf{Q}$ une matrice triangulaire inférieure et

$$V_k = \underbrace{\widetilde{W}_k \oslash ... \oslash \widetilde{W}_k}_{N \text{ fois}}, \tag{3.15}$$

où $\oslash$ désigne le produit de Kronecker modifié et $\widetilde{W}_k = [1, \tilde{w}_k^{(0)}, \tilde{w}_k^{(1)}, \ldots, \tilde{w}_k^{(M-1)}]^\top$. Comme $\widetilde{W}_k$ peut s'écrire comme le produit d'une matrice triangulaire inférieure par $Z_k^w = [1, x_k^w, x_{k-1}^w, ..., x_{k-M+1}^w]^\top$, $V_k$ est aussi le produit d'une matrice triangulaire inférieure par

$$X_k^w = Z_k^w \oslash ... \oslash Z_k^w, \tag{3.16}$$

Ainsi, comme pour les systèmes sans mémoire, on a la relation (3.14), avec $\mathbf{\Gamma_k}$ et $X_k^w$ différents.
Le système adaptatif est ainsi piloté par un vecteur orthonormal, dont la matrice de corrélation est théoriquement la matrice identité. La relation (3.14) permet de comparer théoriquement les performances de notre schéma d'identification avec une identification directe par $X_k$ ou $X_k^w$, le système adaptatif identifiant le système non-linéaire $A$ s'écrivant $A_k^o = \mathbf{\Gamma_k}^\top \psi_k$.

FIGURE 3.4 – Schéma d'identification adaptative de système non-linéaire modélisable par un filtre de Volterra.

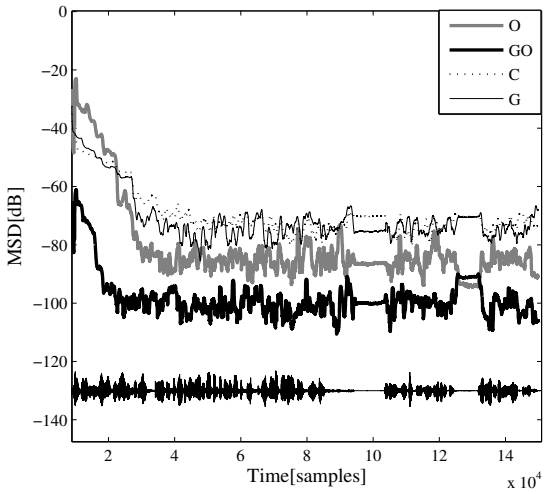



FIGURE 3.5 – Déviation quadratique moyenne pour différents schémas d'identification : identification directe avec le signal original (C) ou le signal gaussianisé (G), identification via le schéma proposé avec le signal original (O) ou le signal gaussianisé (GO). Paramètres de simulation : signal de parole échantillonné à 8 kHz, pas d'adaptation $\mu = 0.1$, ordre de non-linéarité $N = 3$, mémoire $M = 3$, rapport signal à bruit ambiant = 40 dB, rapport signal à bruit de gaussianisation =-18 dB.

Pour une identification par l'algorithme LMS, la rapidité de convergence dépend du conditionnement de la matrice de corrélation du signal d'entrée du modèle adaptatif, $\mathrm{E}[X_k X_k^\top]$ dans le cas d'une identification directe, $\mathrm{E}[U_k U_k^\top]$ avec notre schéma. Cette dernière quantité vaut 1, ce qui assure une convergence optimale. Cependant, le signal n'étant pas stationnaire malgré le blanchiment et l'orthonormalisation, l'algorithme NLMS est préférable, de sorte que la vitesse de convergence est régie par le conditionnement de $\mathrm{E}[U_k U_k^\top / \|U\|^2]$, *a priori* différent de 1 mais que l'on peut raisonnablement supposer meilleur que celui de $\mathrm{E}[X_k X_k^\top / \|X\|^2]$. Nous avons montré que les performances en régime permanent sont d'autant meilleures qu'est faible la quantité $\mathrm{E}[1/\|X_k\|^2]$ dans le cas d'une identification directe ou $\mathrm{E}[\|\mathbf{\Gamma_k} U_k\|^2/\|U_k\|^4]$ avec notre schéma, avec $\mathrm{E}[\|U_k\|^2] \simeq 1$. On peut s'attendre à ce que la seconde valeur soit plus faible et ait des variations plus lisses que la première.

Les simulations confirment ces conjectures issues des calculs théoriques. Le conditionnement est nettement meilleur dans notre schéma, tout en dépassant 1 cependant, ce qui peut s'expliquer par ce que : (i) le signal ne soit pas parfaitement gaussien ; (ii) on considère $\mathrm{E}[U_k U_k^\top / \|U\|^2]$ ; (iii) les sorties du prédicteur ne soient pas parfaitement orthogonales. La figure 3.5 illustre les performances de la méthode proposée, en régime transitoire et en régime permanent. L'identification directe avec un signal gaussianisé et l'identification via le schéma proposé mais sans gaussianisation offrent des performances intermédiaires. Le gain provient donc à la fois de la gaussianisation et du schéma d'identification, lui-même reposant sur la gaussianisation. Nous avons simulé notre système avec un modèle réaliste de haut-parleur inspiré de [43], de mémoire 6 ms et d'ordre de non-linéarité 3 : le rapport signal à erreur calculé sur le signal de sortie est amélioré de 15 dB par rapport à une identification directe.

Une collaboration avec le constructeur d'enceintes Cabasse nous a permis de prolonger cette étude par la caractérisation de haut-parleurs. Les fonctionnements linéaire et non-linéaire d'un haut-parleur sont caractérisés respectivement par sa réponse fréquentielle et par son taux de distorsion harmonique (TDH). Les entrées

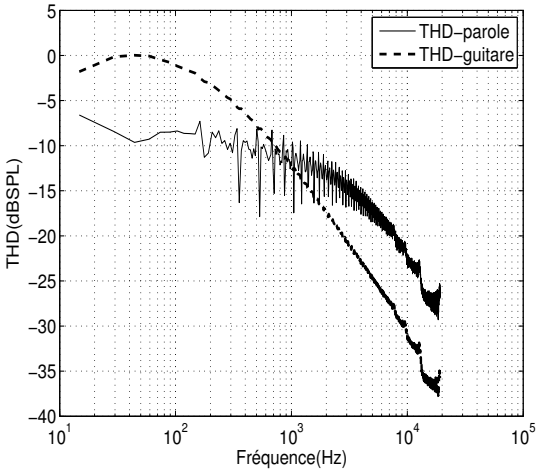



FIGURE 3.6 – Taux de distorsion harmonique de la parole et de la guitare pour un haut-parleur bas-de-gamme, à partir de l'identification du système par la méthode de la figure 3.4.

utilisées pour mesurer ces grandeurs sont respectivement du bruit rose [2] et des sinusoïdes glissantes (*sweeps*) dont la fréquence augmente exponentiellement dans le temps. Le comportement des haut-parleurs dépendant du signal d'entrée, il serait plus pertinent de les caractériser avec les signaux d'usage en entrée. Mais le bruit rose et les sweeps sont plus adaptés aux calculs respectifs de la réponse fréquentielle et du TDH. Comme l'identification d'un haut-parleur dépend de l'entrée mais que la caractérisation d'un système de Volterra fixe est indépendante de l'entrée, nous avons proposé le protocole de caractérisation suivant pour les haut-parleurs :

1. le haut-parleur est modélisé par un filtre de Volterra identifié selon la méthode de la figure 3.4 pour différents types de signaux d'entrée (parole, musique de divers genres), il en résulte un modèle pour chaque entrée ;
2. chaque modèle est excité par un bruit rose pour calculer la réponse fréquentielle et par une sweep pour calculer le TDH, il en résulte une réponse fréquentielle et un TDH pour chaque type d'entrée.

La figure 3.6 montre l'intérêt de cette multi-caractérisation, qui exploite à la fois les résultats de Klippel sur la dépendance des haut-parleurs au signal d'entrée et notre méthode d'identification de système non-linéaire fondée sur la gaussianisation.

### Estimation des paramètres d'un mélange

Les paramètres d'un mélange audio instantané sous-déterminé peuvent être estimés à partir du spectrogramme selon l'approche SCA+ICA proposée dans [26]. Considérant $p$ mélanges de $n$ sources, la méthode consiste à :

1. diviser le plan temps-fréquence en blocs homogènes et traiter chaque bloc par une analyse en composantes indépendantes (ICA), en supposant que seules $p$ sources sur $n$ sont actives dans ce bloc, ce qui fournit une matrice locale de mélange de dimensions $p \times p$ ;
2. appliquer une classification hiérarchique ascendante sur l'ensemble des directions des colonnes des différentes matrices locales de mélange, ce qui permet

2. séquences à longueur maximale filtrées pour avoir un spectre proche de celui de l'audio

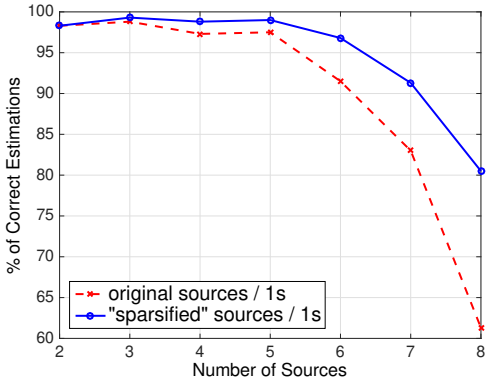



FIGURE 3.7 – Pourcentage d'estimations correctes du nombre de sources dans un mélange de 1 s.

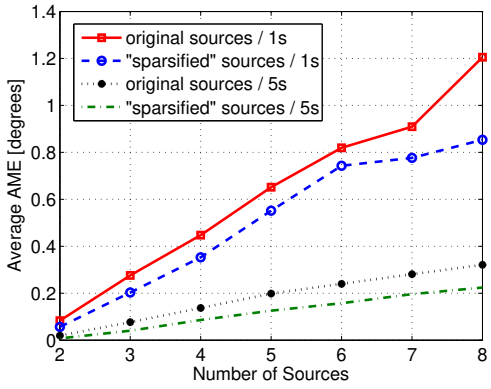



FIGURE 3.8 – Erreur angulaire moyenne d'estimation des colonnes de la matrice de mélange.

de déterminer le nombre de sources $n$ (nombre de classes) et la direction de chaque colonne de la matrice de mélange globale $p \times n$ (centroïdes des classes).

Nous avons appliqué cette méthode [MNR12] sur des mélanges stéréo de 2 à 8 sources de 1 à 5 s, d'une part avec les sources originales et d'autre part avec les sources super-gaussianisées par le reformage global d'histogramme dans le domaine temps-fréquence présenté en 3.1.2. Comme illustré par la figure 3.7, sur une durée de 1 s, la super-gaussianisation réduit d'autant plus le pourcentage d'erreur de comptage des sources que leur nombre est important (le taux d'erreur est inférieur à 2 % dans tous les cas pour une durée de 5 s). De même (voir figure 3.8), la super-gaussianisation réduit l'erreur angulaire d'estimation des colonnes de la matrice de mélange.

Cette application suppose d'avoir accès aux sources originelles ; elle n'est donc envisageable que dans le contexte de séparation de sources informée (ISS) présenté au chapitre précédent.

#### Application du théorème de sous-quantification

Si les conditions du théorème de quantification sont vérifiées, l'histogramme d'un signal peut être parfaitement reconstruit à partir de celui de sa version sous-quantifiée d'un facteur $K$, en appliquant l'équation (3.6). Nous avons montré [HMJ09] que l'erreur de reconstruction est considérablement réduite lorsque le signal a subi, avant sous-quantification, un reformage local d'histogramme (voir sous-section 3.1.2) filtrant passe-bas l'histogramme avec une fréquence de coupure de $1/2K$. Cette erreur n'est cependant pas nulle, ce qui peut s'expliquer par l'utilisation d'histogrammes au lieu de ddp. Le filtrage passe-bas de l'histogramme produit en effet des valeurs décimales, qui doivent être arrondies pour pouvoir constituer un histogramme, ce qui génère des variations haute-fréquence sur les queues de la distribution, de sorte que la fonction caractéristique n'est plus strictement à support borné.

### 3.1.4 Égalisation d'histogramme contrôlée perceptivement

Dans l'utilisation des méthodes d'égalisation d'histogramme présentées en 3.1.2, nous avons montré que, selon le réglage des paramètres de transformation, l'égali-

sation peut être inaudible et améliorer l'histogramme en vue d'un traitement cible. Ainsi, lors de la super-gaussianisation des échantillons temps-fréquence, le choix du facteur de forme cible et la durée du signal conditionnent la qualité du signal modifié ; lors du filtrage passe-bas de l'histogramme, l'audibilité de la transformation du signal dépend de la fréquence de coupure de ce filtrage. Mais ces réglages restent empiriques et ces méthodes d'égalisation ne garantissent pas par elles-mêmes la préservation de la qualité audio. La question du compromis optimal entre cette qualité et l'intensité de la transformation reste donc ouverte. La limitation de l'amplitude du bruit de transformation dans [MMJS+07] et [MMMDL+14] amorce un contrôle perceptif en ce qu'elle limite la puissance de ce bruit. Nous avons prolongé cette piste en formalisant le problème comme suit : considérant un histogramme cible $f_{target}$, **comment transformer un signal $x$ d'histogramme $f_x$ en un signal $z$ d'histogramme $f_z$ tel que la distance $(f_z, f_{target})$ soit minimale, sous la contrainte que $z$ soit perceptivement identique à $x$ ?** Les histogrammes sont soit ceux des échantillons temporels, soit ceux des échantillons temps-fréquence.

Le choix de la distance entre histogrammes dépend de la transformation visée. Pour une transformation globale (HGR), il est préférable d'utiliser une distance impliquant les histogrammes cumulatifs, comme la distance de Kolmogorov-Smirnov (notée $\mathrm{d_{KS}}$ par la suite) : l'intégration sous-jacente de l'histogramme lisse les différences locales et permet de ne tenir compte que de la forme globale de l'histogramme. Au contraire, pour une transformation locale (HLR), une distance comme celle de la variation totale (notée $\mathrm{d_{TV}}$ par la suite) est plus appropriée, car sensible aux différences locales entre les histogrammes.

La formalisation de l'inaudibilité de la transformation de $x$ en $z$ dépend du domaine de transformation : dans le domaine temps-fréquence, la dégradation est une distorsion spectrale variant dans le temps ; dans le domaine temporel, il s'agit d'un bruit additif. Nous avons proposé deux algorithmes adaptés à chacun de ces cas.

### Perceptual HGR dans le domaine temps-fréquence

Nous avons proposé [MNSR14] l'algorithme itératif suivant, fondé sur la comparaison entre $F_{|Z|}$ et $F_{target}$ dans un voisinage de chaque $|Z(m,f)|$. En initialisant $Z$ à $X$ et en parcourant plusieurs fois le spectrogramme,

- si $F_{|Z|} < F_{target}$ sur $I = [|Z(m,f)|_{\mathrm{dB}} - \Delta; |Z(m,f)|_{\mathrm{dB}}[$ , alors on réduit $|Z(m,f)|_{\mathrm{dB}}$ de $\Delta$ ;
- sinon, si $F_{|Z|} > F_{target}$ sur $I = [|Z(m,f)|_{\mathrm{dB}}; |Z(m,f)|_{\mathrm{dB}} + \Delta$, alors on augmente $|Z(m,f)|_{\mathrm{dB}}$ de $\Delta$ ;

de sorte que $|F_{|Z|} - F_{target}|$ décroît sur l'intervalle $I$. Le processus s'arrête quand $\mathrm{d_{KS}}(f_{|Z|}, f_{target})$ a convergé ou que la différence entre $z$ et $x$ est audible.

La décroissance de $\mathrm{d_{KS}}(f_{|Z|}, f_{target})$ est d'autant plus rapide que $\Delta$ est grand, mais une trop grande valeur peut (i) rendre difficile la vérification d'une des deux conditions et empêcher la convergence ; (ii) accentuer la sensibilité de l'algorithme à l'ordre de parcours des échantillons temps-fréquence ; (iii) dégrader plus rapidement la qualité audio.

La transformation équivaut à un filtrage de réponse fréquentielle $|H(m,f)| = |Z(m,f)|/|X(m,f)|$. Les variations de $|H|$ sur les axes temporel et fréquentiel doivent être contrôlées : des variations trop rapides sur l'axe des fréquences produisent un

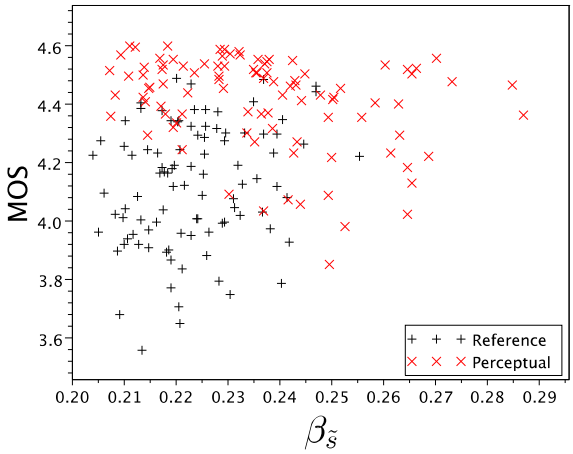



FIGURE 3.9 – Pour 96 signaux de parole, MOS prédit par PESQ *vs* facteur de forme de l'histogramme super-gaussianisé. Comparaison entre l'algorithme de base [MNR12] (« reference » ) et l'algorithme avec contrôle perceptif [MNSR14] (« perceptual » ), les deux ayant pour cible une division par deux du facteur de forme.

effet métallique (voix de robot) ; des variations temporelles trop rapides produisent du bruit musical. Nous avons défini les variations maximales sur les deux axes selon la sensibilité fréquentielle de l'oreille, comme des fonctions croissantes de la fréquence.

L'audibilité de la transformation étant contrôlée tout au long de l'algorithme, nous la mesurons par la distorsion spectrale en échelle Bark (*Bark spectral distance, BSD* [22]), qui a l'avantage d'être peu complexe et bien corrélée avec la distorsion perçue pour les signaux de parole [46]. L'algorithme s'arrête quand la BSD dépasse le seuil d'audibilité ou que l'histogramme cible est atteint.

Cet algorithme a été testé comme celui présenté en 3.1.2 dans le cadre de la super-gaussianisation de signaux de parole, avec un histogramme cible ayant un facteur de forme divisé par deux. Comme précédemment, le facteur de forme est nettement réduit sans atteindre la division par deux. La figure 3.9 montre la différence de compromis qualité / super-gaussianité entre les deux méthodes. Nous avons montré que cette méthode est robuste au codage audio, qui ne change pas le facteur de forme atteint.

### Perceptual HGR / HLR dans le domaine temporel

Le reformage de l'histogramme des échantillons temporels se traduit par un bruit additif. L'égalisation d'histogramme sous contrainte d'inaudibilité de ce bruit peut se traduire par l'optimisation suivante :

$$\begin{cases} \min \mathrm{d}(f_z, f_{target}) \quad \text{sous la contrainte} \ : & (3.17) \\ \gamma_w(m,\nu) < \gamma_{mask}(m,\nu) \quad \forall\, \text{trame}\, m,\ \text{fréquence}\, \nu & (3.18) \end{cases}$$

où d est une distance et, en considérant une analyse par trames quasi-stationnaires, $\gamma_w(m,\nu)$ désigne la densité spectrale de puissance du bruit de transformation $w$ dans la $m^{\text{ème}}$ trame et $\gamma_{mask}(m,\nu)$ le seuil de masquage fréquentiel du signal $x$ dans la $m^{\text{ème}}$ trame.

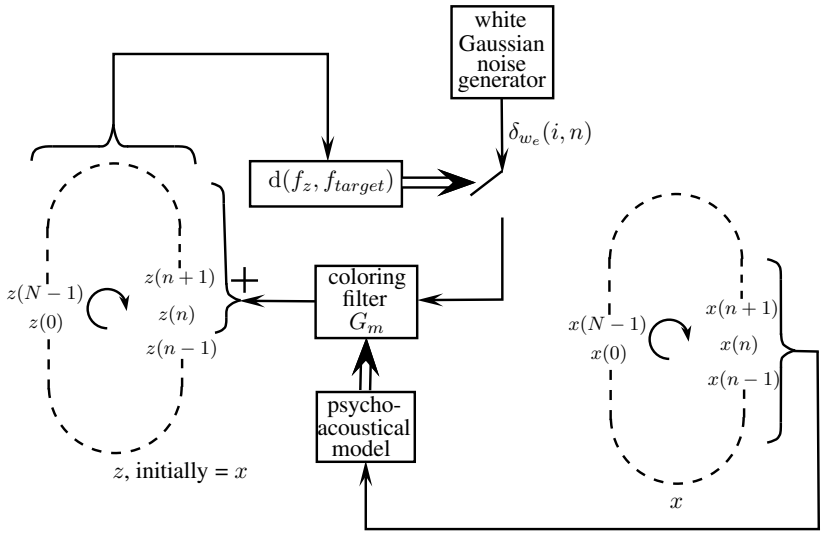



FIGURE 3.10 – Schéma de principe de l'égalisation d'histogramme temporel contrôlée perceptivement.

La difficulté est double : d'une part l'optimisation doit être réalisée sur la globalité du signal alors que la contrainte est locale et différente pour chaque trame ; d'autre part l'histogramme est celui des échantillons temporels alors que la contrainte s'exprime dans le domaine fréquentiel. Nous avons proposé une heuristique *ad hoc* consistant à ajouter itérativement des bruits de faible niveau ayant des spectres parallèles au seuil de masquage (3.18) et contribuant à (3.17). Le schéma de principe présenté dans [MJ18] peut être simplifié selon la figure 3.10.

Au départ, $z = x$. Puis, pour chaque itération $i$ (un cycle sur le signal) et chaque instant discret $n$, on génère une valeur aléatoire $\delta w_e(i,n) \sim \mathcal{N}(0, \sigma_i^2)$. Ajouter à $z$ le filtrage de cette impulsion par le filtre colorant $G_m$ modifierait $z(n) \dots z(n+L)$, avec $L+1$ la longueur de la réponse impulsionnelle $g_m$. Si cette modification réduit $\mathrm{d}(f_z, f_{target})$, alors on fait cet ajout, sinon $z$ reste inchangé. Ainsi, pour $q$ itérations, le bruit ajouté vaut :

$$w = g_m * w_e \tag{3.19}$$

avec

$$w_e(n) = \sum_{i=1}^{q} \delta(i,n) \delta w_e(i,n) \tag{3.20}$$

où $\delta(i,n) = 0$ or 1 selon la décision d'ajouter ou non $\delta w_e(i,n)$ filtré.

Nous avons montré que $w_e$ est blanc, de variance fonction de $q$ et des variances $(\sigma_i)_{1 \le i \le q}$, de sorte que la contrainte (3.18) peut s'exprimer :

$$\begin{cases} |G_m(\nu)|^2 = \gamma_{mask}(m,\nu) & (3.21) \\ \sigma_{w_e} < 1 & (3.22) \end{cases}$$

où la satisfaction de la contrainte (3.22) peut être contrôlée par le choix de $q$ et $(\sigma_i)_{1 \le i \le q}$. L'algorithme s'arrête quand $f_z$ est suffisamment proche de $f_{target}$ ou que $\sigma_{w_e}$ a atteint 1.

Pour le reformage local d'histogramme, la distance entre histogrammes utilisée est la distance de la variation totale. Pour le reformage global, nous utilisons une distance fondée sur les histogrammes cumulés comme expliqué précédemment, mais en

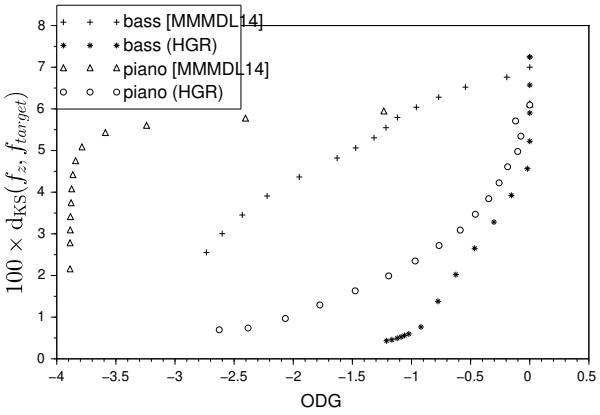



FIGURE 3.11 – Distance de Kolmogorov-Smirnov à l'histogramme cible, $\mathrm{d_{KS}}(f_z, f_{target})$, en fonction de l'ODG prédit par PEAQ, pour des signaux de basse et de piano traités chacun par l'algorithme [MMMDL$^+$14] et par l'HGR contrôlé perceptivement. Pour [MMMDL$^+$14], chaque point correspond à une valeur de $w_{max}$. Pour l'HGR perceptif, chaque point correspond au résultat d'une itération.

distinguant deux niveaux. À la fin de de chaque itération, la distance de Kolmogorov-Smirnov est appropriée ; en revanche, à chaque instant $n$, l'ajout éventuel du bruit filtré modifie localement de l'histogramme, d'une manière qui peut contribuer à rapprocher celui-ci de l'histogramme cible, sans que la distance de Kolmogorov-Smirnov s'en trouve immédiatement modifiée. La décision d'ajout du bruit filtré est donc fondée sur la distance euclidienne entre $F_z$ et $F_{target}$.

Nous avons testé l'HGR contrôlé perceptivement sur différents signaux de musique mono-instrument, pour les super-gaussianiser (en vue d'une séparation de sources dans le domaine temporel). Ces signaux ayant des distributions gaussiennes généralisées, il s'agissait de diviser par deux leur facteur de forme. Selon les instruments, l'algorithme s'arrête soit parce que la distance $\mathrm{d_{KS}}(f_z, f_{target})$ a convergé, soit parce que $\sigma_{w_e}$ a atteint 1, limite de l'audibilité du bruit (sans que la $\mathrm{d_{KS}}(f_z, f_{target})$ soit satisfaisante). Comme illustré par la figure 3.11, le compromis qualité / super-gaussianisation atteint par cette méthode dépasse largement celui de la méthode de base [MMMDL$^+$14] [3].

Nous avons testé l'HLR sur des signaux de parole, pour filtrer passe-bas leur histogramme. Après une décroissance rapide au cours de la première itération, la distance $\mathrm{d_{TV}}(f_z, f_{target})$ converge lentement. Pour une fréquence de coupure de 1/32, $\mathrm{d_{TV}}(f_z, f_{target})$ ne décroît quasiment plus après une centaine d'itérations. À ce stade, le seuil d'audibilité est loin d'être atteint ($\sigma_{w_e} \simeq 0,5$) et l'histogramme est bien filtré passe-bas, comme l'illustre la figure 3.12. L'histogramme cible n'est pas pas parfaitement atteint comme le permettait la méthode [HMJ09], mais l'inaudibilité

3. Un reviewer a malicieusement demandé : *« By reducing the shape parameters (of the GG distribution) by 2 for the $f_{target}$, would not the authors favoring the experimental scenario toward fulfilling more easily the noise inaudibility constraint ? After all, more samples of the modified signal are moved closer to zero. »* La raison était plutôt qu'à ce moment nous nous intéressions plus à la séparation de sources qu'à l'identification de systèmes non-linéaires, qui nécessite au contraire de gaussianiser. Nous avons donc testé les mêmes signaux avec pour chacun un histogramme cible de facteur de forme le double du facteur de forme originel. La convergence est plus rapide que dans l'expérience de réduction du facteur de forme et a lieu avant l'atteinte de la limite d'audibilité. Ces résultats n'ont pas été inclus dans l'article, mais des fichiers audio sont disponibles sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/Histograms/addendum.html`

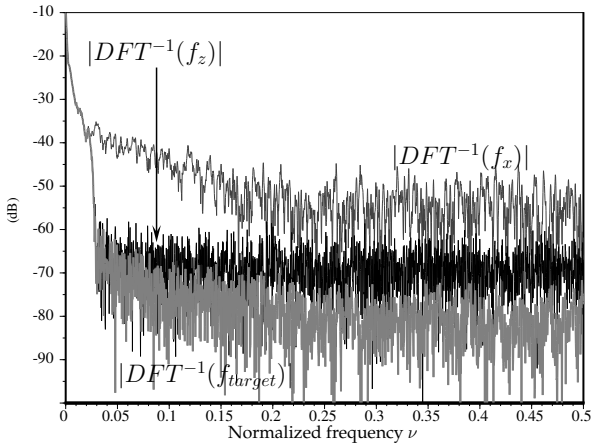



FIGURE 3.12 – Après 100 itérations de l'algorithme, fonctions caractéristiques du signal original ($|\mathrm{DFT}^{-1}(f_x)|$) et du signal à histogramme filtré passe-bas ($|\mathrm{DFT}^{-1}(f_z)|$) avec une fréquence de coupure 1/32, comparés à la fonction caractéristique cible $|\mathrm{DFT}^{-1}(f_{target})|$.

de la transformation est assurée, avec un MOS (estimé par PESQ) entre 4 et 4,5, contre 3,5 à 4 pour la méthode [HMJ09].

### 3.1.5 Applications de l'égalisation d'histogramme contrôlée perceptivement

**Perceptual HGR dans le domaine temps-fréquence → séparation de sources**

Nous considérons la séparation d'un mélange instantané sous-déterminé. Le nombre de sources et la matrice de mélange sont estimés selon la méthode présentée en 3.1.3. La séparation elle-même est réalisée par une méthode NMF multicanal fondée sur une modélisation gaussienne locale [32], fournie par la boîte à outil de séparation de sources audio FASST [33]. Les performances de cette méthode sont améliorées par la donnée de la matrice de mélange précédemment estimée.

Nous avons testé cette séparation [MNSR14] sur des mélanges stéréo de 2 à 6 signaux de parole de 1 et 5 s, en comparant les résultats obtenus avec les signaux super-gaussianisés comme indiqué en 3.1.4 avec les résultats obtenus avec les signaux originaux. Les étapes d'estimation de la matrice de mélange et de séparation ont été réalisées en considérant une parfaite réalisation de l'étape précédente, de manière à étudier l'impact de la super-gaussianisation sur chaque étape séparément.

Les erreurs de comptage des sources et les erreurs angulaires d'estimation des colonnes de la matrice de mélange sont similaires à celles présentées respectivement sur les figures 3.7 et 3.8. Nous avons évalué les performances de la séparation *via* les ratios source à distorsion (SDR), source à interférence (SIR) et source à artefact (SAR) [45]. L'utilisation de sources super-gaussianisées permet un gain d'environ 1,5 dB pour des sources de 5 s, de 1,2 dB pour des sources de 1 s. Pour comparer globalement et perceptivement le processus super-gaussianisation - mélange - séparation avec le processus mélange - séparation, nous avons utilisé les mesures fournies par PEASS [7] en considérant dans les deux cas les signaux originaux comme références. Pour trois des quatre mesures, le gain apporté par la super-gaussianisation est nul ; pour la quatrième, il est significatif à partir de 4 sources. Ces résultats déce-

TABLE 3.1 – Super-gaussianité des sources et qualité de leur mélange, après HGR visant la division par deux du facteur de forme de l'histogramme de chaque source (sauf pour la guitare du mélange B, de manière à ne pas gaussianiser le signal).

| Mélange | instruments | facteur de forme | ODG mélange |
|---|---|---|---|
| A | voix | $1.8 \rightarrow 1.6$ | -0.8 |
| | piano | $2.1 \rightarrow 1.7$ | |
| B | guitare | 3 | -0.4 |
| | claviers | $1.8 \rightarrow 1.3$ | |
| C | voix 1 | $1.1 \rightarrow 1.0$ | -0.6 |
| | voix 2 | $1.0 \rightarrow 0.9$ | |
| D | guitare solo | $1.5 \rightarrow 1.5$ | -0.8 |
| | guitare acoustique | $1.4 \rightarrow 1.1$ | |

vants sont à prendre avec précaution, puisque PEASS n'est pas conçu pour évaluer des dégradations telles que celles introduites par la super-gaussianisation. Tous ces résultats restent valables quand le mélange stéréo subit un codage MP3 à 192 kbit/s, témoignant de la robustesse de l'HGR perceptive à la compression audio.

### Perceptual HGR dans le domaine temporel → séparation de sources

Comme indiqué dans la section 3.1.1, les performances des méthodes de séparation de source fondées sur l'ICA se dégradent si la distribution des sources est trop proche de la gaussianité. Nous considérons le scénario de séparation le plus simple, la séparation d'un mélange stéréo de deux sources, de manière mettre en évidence ce seul facteur de mise en difficulté de l'algorithme de séparation. Nous avons comparé les résultats d'un algorithme de l'état de l'art, FastICA [12], sur des mélanges de sources super-gaussianisées comme décrit en 3.1.4, avec ceux obtenus sur les mélanges des sources originales. Pour chaque mélange, nous avons estimé la super-gaussianité des sources et la qualité du mélange super-gaussianisé (table 3.1), ainsi que les performances de FastICA (figure 3.13). Pour les mélanges A, B et D, la super-gaussianisation améliore nettement la séparation même si le facteur de forme est peu réduit. L'amélioration est plus légère pour le mélange C, qui était déjà bien séparé sans traitement des sources. Les ODG des mélanges et les SxR mesurent séparément la dégradation liée à la super-gaussianisation et l'amélioration de la séparation résultant de celle-ci. Pour mesurer le gain perceptif global (comparaison entre super-gaussianisation - mélange - séparation et mélange - séparation), nous avons comparé les scores fournis par PEASS [7] en prenant comme référence les signaux originaux dans tous les cas. Les résultats sont similaires à ceux de la figure 3.13.

### HLR dans le domaine temporel → application du théorème de quantification

Nous avons sous-quantifié d'un facteur 16 des signaux de parole, soit directement, soit après filtrage passe-bas de leur histogramme avec une fréquence de coupure de 1/32 selon la méthode présentée en 3.1.4. Puis nous avons restauré leur histogramme originel à partir de celui des signaux sous-quantifiés, en appliquant l'équation de reconstruction (3.6) du théorème de sous-quantification. L'erreur de reconstruction, définie comme la distance de la variation totale entre les histogrammes restauré et originel, est de l'ordre de $10^{-2}$ avec égalisation d'histogramme (HLR), contre $10^{-1}$ sans

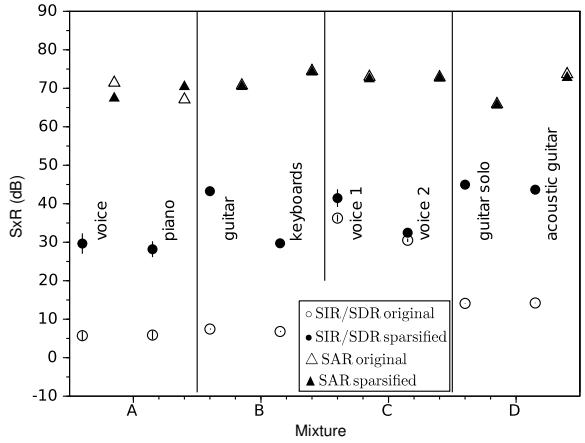



FIGURE 3.13 – Performances de la séparation de sources par FastICA avec et sans super-gaussianisation des sources, pour les 4 mélanges A, B, C et D de la table 3.1. Les barres verticales représentent les intervalles de confiance. SIR = Source to Interference Ratio ; SDR = Source to Distortion Ratio ; SAR = Source to Artifact Ratio.

égalisation. Elle est évidemment plus faible avec la méthode d'égalisation [HMJ09] (de l'ordre de $10^{-3}$), au prix d'une dégradation de la qualité comme on l'a vu précédemment.

Sur l'égalisation d'histogramme contrôlée perceptivement et ses applications, des démonstrations audio et vidéo, ainsi que les codes Scilab des algorithmes, sont disponibles sur `http://up5.fr/SoundHistograms`

### 3.1.6 Discussion sur l'égalisation d'histogramme

L'égalisation d'histogramme s'applique à des domaines variés, avec deux contraintes communes.

D'une part les séquences traitées sont généralement longues (1 s ou plus), de sorte que les méthodes proposées sont adaptées à un traitement hors-ligne du signal audio, même si l'application ultérieure (identification de système, séparation de sources...) peut se faire en temps réel. Envisager des chaînes de traitement entièrement au fil de l'eau suppose soit d'étudier de manière plus systématique l'effet de la durée du signal sur les traitements proposées (en visant des mémoires tampon de l'ordre de la dizaine de ms pour la construction des histogrammes), soit d'adapter les chaînes de traitement proposées à une modification en continu de l'histogramme, au fur et à mesure de l'arrivée de nouveaux échantillons.

D'autre part, notre proposition suppose l'accès aux signaux originaux. Dans quelle mesure cela limite-t-il sa portée ?

— L'accès aux signaux originaux est inhérent à l'identification de systèmes non-linéaires, pour laquelle nous avons une solution performante pour un problème pratique général.

— L'égalisation d'histogramme pour l'application du théorème de sous-quantification présente d'abord un intérêt théorique : elle constitue, à l'égard de ce théorème, l'équivalent du filtre anti-repliement pour le théorème de Shannon. D'un point

de vue plus pratique, ces travaux permettent de viser, à long terme, la possibilité de rendre des signaux audio « déquantifiables » , en adaptant notre solution aux histogrammes multidimensionnels représentatifs de la densité de probabilité jointe d'une séquence de plusieurs échantillons.

— Pour la séparation de sources, l'hypothèse d'accès aux signaux originaux nous limite au contexte applicatif très spécifique de la séparation de sources informée. Les très bonnes performances de l'ICA dans le domaine temporel sont à relativiser, puisqu'en pratique l'ICA peut avantageusement se faire dans le domaine temps-fréquence sans pré-traitement. Cet exemple avait avant tout une valeur démonstrative de l'idée de dopage des signaux audio. Dans le domaine temps-fréquence, le gain apporté par notre méthode à la séparation par SCA (1 à 2 dB de SxR) est modeste comparé à celui de l'état de l'art de la même époque (5 à 20 dB) mais avec des avantages notables : l'absence de canal auxiliaire (5 à 20 kbit/s par source dans l'état de l'art), la robustesse à la compression audio et l'indépendance à la matrice de mélange. Ce gain semble cependant insuffisant dans un contexte informé. Nous verrons dans la section suivante comment le pré-traitement des sources peut mieux prendre en compte les hypothèses de la SCA.

## 3.2 Parcimonisation jointe

**Collaborations :**

— João-Marcos Romano et Everton Nadalin (DSPcom, Unicamp)

— Ricardo Suyama (CECS, Universidade Federal do ABC)

**Projets :** Compest et ROAPI (2014-)

Dans la section précédente, nous avons considéré la parcimonisation des sources audio au sens d'une super-gaussianisation de chaque source indépendamment des autres. Si cette approche permet d'améliorer les performances de la séparation de sources par SCA, elle ne satisfait qu'imparfaitement aux hypothèses idéales de celle-ci, à savoir la parcimonie jointe des sources, c'est-à-dire le fait que dans un certain espace de représentation, les sources aient un grand nombre de coefficients nuls et ne se recouvrent pas. Nous considérons ici une représentation temps-fréquence, telle que classiquement utilisée en séparation de sources audio, et un mélange stéréo instantané de plusieurs sources audio.

Une parcimonisation au sens $\ell_0$ a été proposée par Pinel *et al.* [35], qui perfectionne celle de Balazs *et al.* [1]. Pour chaque trame temporelle, elle consiste à mettre à zéro les composantes fréquentielles masquées par les autres composantes fréquentielles de la trame. Cette méthode permet d'atteindre environ 75 % de zéros par source audio sans distorsion audible, pour des signaux échantillonnés à 44,1 kHz. Comme évoqué au chapitre 2, cette parcimonisation augmente la probabilité qu'à un instant et une fréquence donnés, au maximum deux sources soient actives dans un mélange stéréo, ce qui permet de séparer le mélange par une ICA locale $2 \times 2$ sachant la matrice globale de mélange, en supposant transmises cette dernière et la matrice d'activité de chaque source [34].

Ainsi, la séparation de source informée proposée par Pinel *et al.* relève à la fois du tatouage dopant, puisqu'elle modifie le signal original pour faciliter la séparation, et

du tatouage réflexif, puisqu'elle nécessite la transmission (par tatouage) d'une information auxiliaire décrivant le mélange. Dans le chapitre 2, nous avons montré qu'il était possible de réduire le débit de cette information auxiliaire en ne transmettant que ce que la SCA n'est pas capable d'estimer, c'est-à-dire l'erreur commise par la SCA sur les matrices de dominance. Cette solution se heurtait à deux difficultés : (i) la matrice de dominance calculée à partir des signaux originaux n'est plus valable pour le mélange codé-décodé ; (ii) la matrice erreur est insuffisamment creuse pour que le débit d'information auxiliaire soit compatible avec une transmission de cette information par tatouage. Nous avons provisoirement écarté la première difficulté pour nous concentrer sur la seconde.

Il s'agit donc d'améliorer le dopage du signal de manière à réduire (voire supprimer) l'information auxiliaire nécessaire à la séparation de sources. Notons que la parcimonisation est réalisée jusqu'à présent sur chaque source indépendamment des autres alors que ce n'est qu'un moyen d'atteindre la parcimonie jointe, qui n'est elle-même qu'un moyen d'améliorer la SCA. **Nous proposons donc, pour chaque couple temps-fréquence où la SCA échoue à identifier les deux sources dominantes, de modifier légèrement celles-ci de telle sorte que l'identification réussisse.** Le problème peut être formalisé comme suit.

Soit $\mathcal{C}$ l'ensemble de toutes les paires d'entiers de $[\![1;p]\!]$, avec $p$ le nombre de sources. Dans le cas où la SCA échoue à trouver les sources dominantes $\Lambda \in \mathcal{C}$, on ajoute au vecteur des sources $S$ un bruit $\varepsilon$ dont toutes les composantes sont nulles sauf $\varepsilon_{\Lambda(1)}$ et $\varepsilon_{\Lambda(2)}$. L'estimation de $S+\varepsilon$ sous l'hypothèse que les sources $\Lambda$ soient dominantes est donnée par :

$$\hat{S}_\varepsilon^{(\Lambda)} = A_\Lambda^{-1} A(S+\varepsilon) \tag{3.23}$$

où $A$ est la matrice de mélange et $A_\Lambda$ est $A$ restreinte aux colonnes $\Lambda$.

**Notre but est de trouver le bruit $\varepsilon$ maximisant le ratio masque à bruit (MNR) dans le mélange tout en assurant que la SCA va identifier $\Lambda$ comme les sources dominantes.** Ce que l'on peut exprimer, sous l'hypothèse d'une SCA locale telle que décrite dans [5], section 10.6 :

$$\begin{cases} \min \max \left( \frac{|(A\varepsilon)_1|}{M_1}, \frac{|(A\varepsilon)_2|}{M_2} \right) \\ \text{s.c.} \quad \|\hat{S}_\varepsilon^{(\Lambda)}\|_\tau < \|\hat{S}_\varepsilon^{(\Lambda')}\|_\tau \quad \forall\, \Lambda' \in \mathcal{C} \setminus \{\Lambda\} \end{cases} \tag{3.24}$$

avec $M_1$ and $M_2$ les seuils respectifs de masquage du mélange sur les canaux gauche et droite. Ce problème d'optimisation n'a pas toujours une solution et celle-ci, lorsqu'elle existe, peut correspondre à un bruit $\varepsilon$ audible. Dans nos expériences, l'audibilité est assurée pour un rapport masque à bruit minimal (première partie de l'équation (3.24)) supérieur à -5 dB.

Nous avons testé cette méthode de forçage des sources sur un mélange stéréo de trois instruments, guitare-basse-violoncelle. On obtient une solution acceptable (MNR $> -5$ dB) pour 60 % des couples temps-fréquence nécessitant une modification (c'est-à-dire pour lesquels la SCA échouait à identifier les sources dominantes). Ce forçage réduit la proportion de 1 dans la matrice erreur à transmettre (différence entre matrice de dominance et matrice de dominance estimée par SCA) à 1 % ou moins, de sorte que l'information auxiliaire pour aider la SCA peut être transmise à un débit inférieur à 1 kbit/s par source.

Le mélange a été séparé par SCA en considérant 4 conditions :

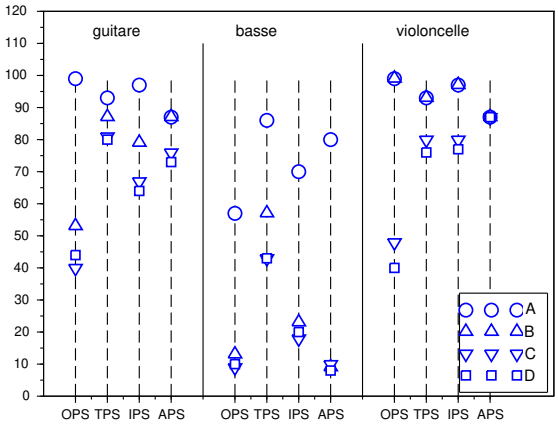



FIGURE 3.14 – Pour un mélange stéréo guitare-basse-violoncelle séparé par SCA dans 4 conditions A/B/C/D, évaluation de la séparation *via* les Overall/Target/Interference/Artifact related Perceptual Scores estimés par PEASS. Les signaux de référence sont les signaux parcimonisés forcés pour les conditions A et B, les signaux parcimonisées non-forcés pour la condition C et les signaux originaux pour la condition D.

(A) Sources parcimonisées forcées, information auxiliaire = matrice de mélange + matrices erreur des matrices de dominance ;

(B) Sources parcimonisées forcées, information auxiliaire = matrice de mélange ;

(C) Sources parcimonisées non-forcées, information auxiliaire = matrice de mélange ;

(D) Sources originales, information auxiliaire = matrice de mélange

La figure 3.14 indique les mesures perceptives de qualité de la séparation fournies par PEASS [7]. Les fichiers audio correspondants sont disponibles sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/HDR/`.

Ces résultats montrent qu'il est possible d'obtenir une très bonne qualité de séparation en combinant un tatouage dopant et un tatouage réflexif d'un débit nettement inférieur à ceux des travaux de référence (1 kbit/s par source contre 5 à 20). Le simple tatouage dopant (parcimonisation + forçage, condition B) permet aussi une nette amélioration par rapport à la SCA effectuée sur le mélange non-traité. Notons toutefois que le peu d'erreurs résiduelles dans l'estimation des sources dominantes par la SCA a un impact fort sur la qualité de séparation. La méthode proposée présente l'intérêt d'utiliser une méthode de SCA très simple en réception, la complexité étant reportée dans le traitement des sources avant mélange.

Ces travaux se poursuivent autour des questions suivantes :

— Si l'on force les sources de manière à améliorer la SCA, la parcimonisation $\ell_0$ préliminaire, qui introduit une dégradation supplémentaire, est-elle nécessaire ?

— Comment réduire encore le débit d'information auxiliaire ?

— Comment prendre en compte le codage-décodage du mélange ?

La pertinence de la poursuite de ces travaux se pose néanmoins, eu égard au positionnement de « niche » de leur seule application pratique imaginée, la séparation de sources informée avec accès aux sources avant mélange. Comme évoqué dans

la conclusion du chapitre 2, cette piste semble promise à l'obsolescence par les formats de codage de type SAOC, conçus pour permettre la séparation d'un mélange stéréo. De fait, les chercheurs impliqués dans l'ISS, qui proposaient vers 2010 des méthodes fondées sur la transmission par un canal auxiliaire (tatouage par exemple) d'informations volumineuses sur les sources et sur le mélange, se sont tournés vers une information de la séparation par des données externes de haut-niveau facilement accessibles, ne nécessitant ni accès aux sources avant mélange, ni canal auxiliaire de transmission. Ces données peuvent être les paroles d'une chanson [38], la partition musicale d'un morceau [8] ou l'activité cérébrale de l'auditeur se concentrant sur un instrument [3].

## 3.3 Quantification auto-correctrice

**Collaborations :**

— Mamadou Mboup (CRESTIC, Université de Reims Champagne Ardenne)

— Monia Turki (U2S, ENIT)

**Encadrement :** thèse de Fatimetou El Jili (Université de Reims Champagne Ardenne, 2015-2018), co-encadrée avec Mamadou Mboup

**Projet :** ICityForAll (2015-)

**Publications :**

[EJMM17] F. El-Jili, G. Mahé, and M. Mboup. A robust signal quantization system based on error correcting codes. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2561–2565, Aug 2017.

[MEJM] G. Mahé, F. El-Jili, and M. Mboup. A robust signal quantization based on error correcting codes. soumis au *Journal on Advances on Signal Processing*.

Le projet ICityForAll visait notamment à aider les personnes presbyacousiques à détecter, identifier et localiser les alarmes de véhicules prioritaires lorsqu'elles conduisent, la presbyacousie réduisant ces capacités. La détection et l'identification étaient l'objet de la thèse de Fatimetou El Jili, dirigée par Mamadou Mboup à l'Université de Reims et initialement envisagée en co-tutelle avec l'ENIT sous la co-direction de Monia Turki. À la suite de péripéties administratives qui ont empêché la co-tutelle, Mamadou Mboup m'a proposé de co-encadrer cette thèse, avec une marge de liberté sur les orientations plus importante que celle permise initialement par le projet. J'ai proposé de **rendre les signaux audio (dont les alarmes) robustes au bruit ambiant en les recomposant comme des suites de combinaisons de codes correcteurs d'erreur, ce qui permettrait, par les techniques de codage de canal, d'annuler le bruit et de faciliter la détection et l'identification.** La thèse de Fatimetou s'est donc initialement concentrée sur la détection des alarmes [25], puis a développé cette idée de signaux robustes, pour enfin revenir à l'application initiale.

### 3.3.1 Principes

On considère la transmission d'un signal (audio par exemple) à travers un canal bruité. Notre objectif est de rendre le signal robuste au bruit du canal, sans changer d'espace de représentation.

Considérons la décomposition binaire signe + valeur absolue sur $L+1$ bits d'un bloc de $n$ échantillons (temporels ou autre, selon l'espace de représentation) du signal, $x = [x_0, x_1, \ldots, x_{n-1}]$ :

$$x = XS_x, \quad \text{avec } X = \sum_{i=0}^{L-1} 2^i X_i, \tag{3.25}$$

où $X_i \in \mathbb{F}_2^n = (\mathbb{Z}/2\mathbb{Z})^n$ pour $0 \leq i \leq L-1$ et $S_x$ est une matrice diagonale contenant les signes de $x$, c'est-à-dire $(S_x)_{ii} = \text{signe}(x_{i-1}), i = 1 \ldots L$.

Nous proposons de remplacer, pour chaque poids de bit $i$, le vecteur $X_i$ par un mot de code $C_i$ de même longueur généré par un codeur correcteur d'erreur de dimension $k_i < n$, de telle sorte que le nouveau vecteur $\tilde{x}$ :

$$\tilde{x} = \tilde{X}S_x, \quad \text{avec } \tilde{X} = \sum_{i=0}^{L-1} 2^i C_i, \tag{3.26}$$

soit aussi proche que possible de $x$. Nous discuterons plus loin de la définition de « proche » .

Supposons maintenant que le signal ainsi re-quantifié $\tilde{x}$ soit transmis à travers un canal équivalent à un bruit additif. Le signal reçu $y$ peut s'écrire comme dans l'équation (3.25) :

$$y = YS_y, \quad \text{avec} \quad Y = \sum_{i=0}^{L-1} 2^i Y_i, \tag{3.27}$$

Les mots de code originaux $C_i$, peuvent être retrouvés à partir des vecteurs $Y_i$ par des méthodes classiques de décodage de canal, ce qui permet de restaurer $\tilde{x}$, en supposant que $S_y = S_x$ et que le pouvoir de correction des codes utilisés soit suffisant eu égard à la quantité d'erreurs binaires provoquées par le bruit du canal.

Notre proposition diffère de l'approche classique codage de source + codage de canal en ce que le résultat final n'est pas un flux binaire. Elle diffère aussi du codage conjoint source-canal, qui convertit le signal dans l'espace de modulation. Ici, le signal reste dans son espace de représentation initial. Nous nous sommes néanmoins inspiré à la fois du codage de canal et du codage de source. Comme dans ce dernier, le signal est approché par une combinaison réduite de vecteurs d'un dictionnaire, avec la différence suivante : alors que le codage de source utilise un dictionnaire construit à partir des données de manière à minimiser un critère d'erreur quadratique moyenne, les dictionnaires que nous utilisons maximisent la distance entre leurs éléments, de manière à les rendre robustes au bruit. Notre quantification diffère aussi de la quantification vectorielle utilisée en codage de source : au lieu d'être une application de $\mathbb{R}^n$ vers un sous-ensemble de $\mathbb{R}^n$, elle consiste en $L$ applications de $\mathbb{F}_2^n$ vers des sous-espaces de dimensions $k_i$ de $\mathbb{F}_2^n$, qui peuvent être indépendantes les unes des autres.

Le choix des codeurs sera discuté en détail plus loin. On peut toutefois intuitivement prévoir que les niveaux de bit les plus faibles seront les plus sensibles au

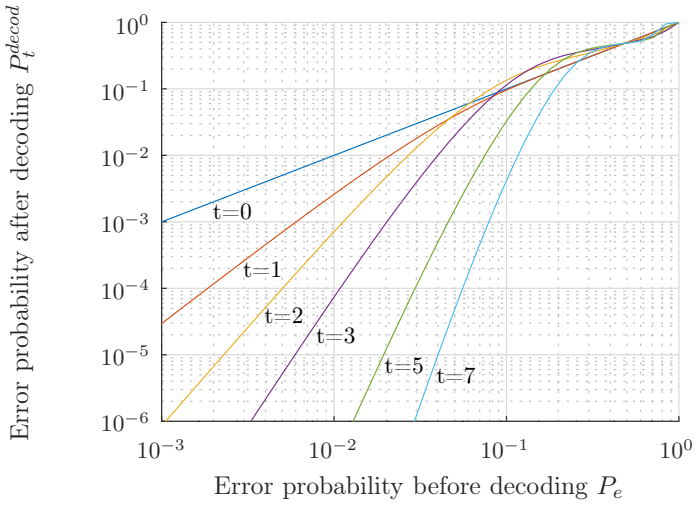



FIGURE 3.15 – Pour des codes BCH de longueur 31 avec différents pouvoirs de correction $t$, probabilité d'erreur après décodage en fonction de la probabilité d'erreur avant décodage.

bruit et nécessiteront donc des pouvoirs de correction plus importants. De ce point de vue, la quantification proposée peut être comparée à la modulation codée en blocs (BCM [13]), qui optimise conjointement le codage et la modulation, à la différence suivante près : alors que la BCM transforme des mots de $k_i$ bits en mots de $n$ bits, notre quantification conserve aux vecteurs binaires leur dimension, si bien qu'elle s'apparente plus au décodage de canal qu'au codage.

### 3.3.2 Choix des codeurs

Pour des mots de code de longueur $n$ générés par un codeur de pouvoir de correction $t$ et transmis par un canal binaire symétrique avec une probabilité d'erreur $P_e$, la probabilité d'erreur binaire après décodage est donnée par :

$$P_t^{decod} = \frac{1}{n} \sum_{j=t+1}^{n} N_{decod}(j,n) \binom{n}{j} P_e^j (1-P_e)^{n-j}, \tag{3.28}$$

où $N_{decod}(j,n)$ est le nombre moyen d'erreurs après décodage quand $j$ des $n$ bits d'un code sont erronés. La figure 3.15 illustre cette relation pour un codeur BCH [2] de longueur 31.

Ainsi, connaissant la probabilité d'erreur pour un niveau de bit donné dans la représentation binaire du signal, on peut fixer le pouvoir de correction adéquat pour les codes évoqués dans l'équation 3.26, selon la probabilité d'erreur visée après décodage.

La probabilité d'erreur avant décodage peut être calculée à partir des densités de probabilité du signal et du bruit du canal. Considérons l'ajout d'un échantillon de signal $s$ et d'un échantillon de bruit $v$, en représentation signe + valeur absolue sur $L+1$ bits, c'est à dire $\pm s_{L-1} \dots s_0$ et $\pm v_{L-1} \dots v_0$, avec $s_i$ et $v_i \in \{0,1\}$ pour $0 \leq i \leq L-1$. Si leurs signes sont différents, on considère la soustraction $||s| - |v||$. L'erreur sur le poids $i$ provient à la fois de la valeur du $i^{\text{ème}}$ bit du bruit et de la retenue issue des poids inférieurs. On note $c_i$ la retenue au rang $i$, résultant de l'addition ou de la soustraction élémentaire au rang $i-1$ pour $i \geq 1$ ($c_0 = 0$). Une erreur se produit sur le $i^{\text{ème}}$ bit si $v_i \neq c_i$.

Nous avons montré que la probabilité d'erreur binaire sur le $i^{\text{ème}}$ bit $s_i$ peut s'exprimer :

$$P_e(i) = \beta_i + \frac{1}{2}(\rho_i^+ + \rho_i^-)(1 - 2\beta_i), \tag{3.29}$$

où

$$\beta_i \triangleq \Pr(v_i = 1) = \Pr(2^i \leq |v| < 2^{i+1}) + \frac{1}{2}\Pr(|v| \geq 2^{i+1}) \tag{3.30}$$

$$\rho_i^+ \triangleq \Pr(c_i = 1 \mid \text{signe}(v) = \text{signe}(s)) \tag{3.31}$$

$$\rho_i^- \triangleq \Pr(c_i = 1 \mid \text{signe}(v) \neq \text{signe}(s)) \tag{3.32}$$

Les probabilités conditionnelles $\rho_i^+$ et $\rho_i^-$ valent 0 pour $i = 0$ et, pour $i \geq 0$,

$$\rho_{i+1}^+ = \rho_i^+(\alpha_i + \beta_i - 2\alpha_i\beta_i) + \alpha_i\beta_i, \tag{3.33}$$

avec $\alpha_i \triangleq \Pr(s_i = 1)$, et $\rho_{i+1}^-$ s'exprime par une formule de récurrence faisant intervenir les quantités suivantes :

$$\Pr\left(b \geq 2^i \mid a \geq 2^j, \text{signe}(v) \neq \text{signe}(s)\right) \tag{3.34}$$

$$\Pr\left(b_i = 1 \mid b \geq 2^i, \text{signe}(v) \neq \text{signe}(s)\right) \tag{3.35}$$

$$\Pr\left(a_i = 1 \mid a \geq 2^i, \text{signe}(v) \neq \text{signe}(s)\right) \tag{3.36}$$

où $a \triangleq \max(|v|, |s|)$ et $b \triangleq \min(|v|, |s|)$.

La complexité du calcul de $\rho_i^-$ est liée aux hypothèses d'indépendance possibles, notamment au fait que les événements $\{c_i = 1\}$ et $\{s_i = v_i\}$ sont indépendants sachant $\{\text{signe}(v) \neq \text{signe}(s), a \geq 2^i\}$, mais pas sachant simplement $\{\text{signe}(v) \neq \text{signe}(s)\}$, ce qui n'est pas intuitif.

La probabilité d'erreur par poids de bit est représentée sur la figure 3.16 pour un signal et un bruit de distributions laplaciennes. La légère différence entre la courbe théorique et la courbe empirique vient de ce que les hypothèses d'indépendance sur lesquelles reposent la preuve de la formule (3.29) ne sont pas parfaitement vérifiées (notamment l'indépendance entre $a_i$ et $b_i$).

**Ainsi, connaissant les densités de probabilité du signal et du bruit, à partir des formules (3.29) et (3.28), on peut déterminer, pour chaque niveau de bit, le pouvoir de correction nécessaire des mots de code utilisés.**

### 3.3.3 Codage et décodage

Recomposer le signal comme une combinaison de mots de codes selon l'équation (3.26) peut se faire simplement en remplaçant chaque vecteur $X_i$ par le mot de code $C_i$ minimisant la distance de Hamming [EJMM17], comme dans un décodage de canal classique. Cette solution a l'avantage d'avoir une complexité faible, mais la distance de Hamming par niveau de bit est peu représentative de la distorsion créé par cette requantification. Nous avons donc cherché à minimiser l'erreur quadratique $\|x - \tilde{x}\|_2^2$, que l'on peut aussi écrire $\|X - \tilde{X}\|_2^2$.

Rechercher les $L$ codes $(C_0, \ldots, C_{L-1})$ tels que le vecteur requantifié $\tilde{x}$ défini par (3.26) minimise $\|x - \tilde{x}\|_2^2$ est un problème d'optimisation complexe. Nous avons contourné cette complexité par un algorithme classique de matching pursuit [6,

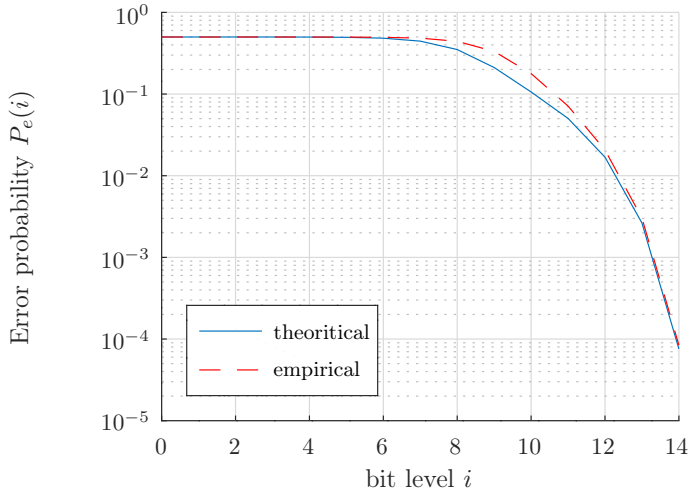



FIGURE 3.16 – Probabilité d'erreur par poids de bit pour un signal laplacien centré d'écart-type $0.1 \times 2^{15}$ corrompu par un bruit laplacien centré d'écart-type $0.01 \times 2^{15}$. Les probabilités empiriques sont obtenues par une simulation sur $10^6$ échantillons.

11] que nous avons adapté à notre cas : voir algorithme 1. Maximiser simplement $\langle C|R_{i-1}\rangle$ favoriserait les codes ayant des poids de Hamming élevés, ce qui augmenterait l'erreur de quantification. C'est pourquoi ce produit scalaire est pénalisé par le poids de Hamming $w_C$.

**Algorithme 1** : Matching pursuit modifiée pour la quantification. $w_C$ désigne le poids de Hamming de $C$

**Entrées** : le vecteur $X$ à requantifier
**Sorties** : le vecteur requantifié $\tilde{X}$
$\tilde{X}_0 \leftarrow 0_{1\times n}$ ; $R_0 \leftarrow X$
**pour** $i \leftarrow 1$ *à* $L$ **faire**
  $C_{L-i} \leftarrow \arg\max_{C\in D_{L-i}} \langle C|R_{i-1}\rangle - \lambda 2^{L-i} w_C$
  $R_i \leftarrow R_{i-1} - 2^{L-i} C_{L-i}$
  $\tilde{X}_i \leftarrow \tilde{X}_{i-1} + 2^{L-i} C_{L-i}$
**fin**
**retourner** $\tilde{X} = \tilde{X}_L$

La valeur optimale du facteur de pénalisation $\lambda$ est difficile à trouver théoriquement mais peut être obtenue empiriquement pour une combinaison donnée de dictionnaires de codes, en quantifiant selon l'algorithme 1 un grand nombre de vecteurs aléatoires et en mesurant l'erreur quadratique finale pour différentes valeurs de $\lambda$. Pour différentes lois de tirage des vecteurs aléatoires, l'optimum est atteint pour $\lambda = 1$ avec des codes BCH. La valeur optimale de $\lambda$ est probablement différente pour des turbo-codes, dont le poids de Hamming est maximisé.

Le décodage du signal requantifié puis bruité peut se faire par le même algorithme, mais cette solution n'exploite pas (ou du moins pas explicitement) le pouvoir de correction des codes correcteurs. Nous décomposons donc chaque bloc $y$ du signal reçu selon l'équation (3.27) et nous retrouvons les mots de codes originels par un décodage de canal classique, avec une complexité faible. En reprenant le formalisme

de la sous-section 3.3.1, le bloc estimé $\hat{x}$ s'écrit :

$$\hat{x} = \hat{X} S_y, \quad \text{avec } \hat{X} = \sum_{i=0}^{L-1} 2^i \hat{X}_i, \tag{3.37}$$

tel que

$$\forall\, 0 \leq i \leq L-1, \quad \hat{X}_i = \mathsf{decod}(Y_i), \tag{3.38}$$

où $\mathsf{decod}(\cdot)$ désigne une fonction de décodage fournissant le mot de code le plus proche au sens de la distance de Hamming.

Seule la valeur absolue de chaque échantillon peut être corrigée de cette manière, puisque l'on conserve le signe de l'échantillon bruité, ce qui peut provoquer des pics d'erreur si l'échantillon de bruit est de signe opposé à celui du signal et plus grand en valeur absolue. C'est pourquoi nous corrigeons le signe à partir d'une prédiction linéaire adaptative fondée sur les échantillons voisins : si l'opposé de l'échantillon corrigé est plus proche du prédicteur que l'échantillon lui-même, alors on remplace ce dernier par son opposé.

### 3.3.4 Applications

Un compromis doit être trouvé entre le pouvoir de correction des codes utilisés et le bruit de codage, de sorte que le signal après codage-bruitage-décodage soit moins bruité que le signal bruité. Le codage par l'algorithme 1 permet de réduire le rapport signal à bruit (RSB) de codage par rapport à la méthode initiale [EJMM17] (réduction de 4 dB pour un signal de distribution uniforme perturbé par un bruit gaussien), mais il semble difficile, d'après les simulations réalisées dans la thèse de Fatimetou El Jili, d'obtenir un RSB global pour la chaîne codage-bruitage-décodage meilleur que celui résultant du bruit seul. La technique proposée est donc appropriée à deux cas pratiques.

Le premier est celui de signaux artificiels construits à partir des dictionnaires de mots de code, pour lesquels la notion de fidélité à un signal original ne s'applique pas. C'est à peu près le cas des alarmes étudiées dans la thèse de Fatimetou, pour lesquelles l'important est de respecter la signature temps-fréquence de l'alarme.

Le second cas concerne les sons, les images et les vidéos, et consiste en un canal introduisant une dégradation perceptivement plus gênante que le bruit de codage, bien que moins forte en termes de rapport signal à bruit.

#### Débruitage et détection d'alarmes de véhicules prioritaires

Les alarmes des véhicules prioritaires sont caractérisées par une *signature temps-fréquence* : le signal consiste en l'alternance de deux signaux harmoniques de fréquences fondamentales respectives $F_1$ (autour de 400 Hz) et $F_2$ (autour de 600 Hz) pour des durées respectives $T_1$ et $T_2$, de 0,5 à 1 s. Ces paramètres dépendent du service et du pays.

La méthode de requantification précédemment décrite a été appliquée à ces signaux. Selon le RSB du canal pour lequel les codeurs sont choisis, il en résulte un RSB de codage de 7 à 9 dB avec l'algorithme 1 appliqué dans le domaine temporel.

Si ce RSB peut sembler faible, la signature temps-fréquence de l'alarme est parfaitement conservée et le spectrogramme de puissance est proche du spectrogramme original, avec cependant plus de puissance dans les hautes fréquences. En revanche, l'application de la méthode à une représentation temps-fréquence du signal noie le spectrogramme dans le bruit et rend le signal méconnaissable.

Lorsque ces signaux sont transmis à travers un canal bruité, le RSB calculé entre le signal décodé et le signal codé est égal à celui du canal plus 6 dB : le pouvoir de correction des codeurs choisis est tel que le décodage ne corrige pas toutes les erreurs provoquées par le bruit du canal, mais corrige suffisamment pour réduire de 6 dB celui-ci. Nous avons étudié dans quelle mesure cette réduction de bruit améliore la détection des alarmes.

Un algorithme de détection d'alarme a été proposé dans [25], fondé sur l'analyse des signatures fréquentielle et temporelle. La signature fréquentielle est calculée par intégration du spectrogramme dans le temps ; l'alarme recherchée est détectée par détection des pics fréquentiels dans le voisinage des harmoniques de $F_1$ et $F_2$, avec une marge $\epsilon$ tenant compte de l'effet Doppler. La signature temporelle est calculée par intégration du spectrogramme sur l'axe des fréquences ; l'alarme recherchée est détectée par détection des discontinuités à des instants espacés alternativement de $T_1$ et $T_2$, avec une marge $\epsilon'$. Dans la thèse de Fatimetou, nous avons étudié les courbes ROC paramétrées par $\epsilon, \epsilon'$ et le nombre $\mu$ d'harmoniques requis pour la détection de la signature fréquentielle, pour différents rapports signal à bruit, en considérant un bruit de trafic. On peut retenir de cette étude que

— pour tous les RSB, la détection est meilleure en utilisant la signature fréquentielle seule qu'en utilisant conjointement les signatures fréquentielle et temporelle ;

— le nombre optimal d'harmoniques (paramètre $\mu$) augmente avec le niveau de bruit ;

— la méthode est efficace pour la détection mais peu adaptée à la reconnaissance d'alarme, car les fréquences fondamentales des différentes alarmes ont des différences du même ordre de grandeur que la marge $\epsilon$ nécessaire pour que la détection soit robuste à l'effet Doppler.

La figure 3.17 compare les performances de cet algorithme de détection sur une sirène de pompiers en trois versions : originale ; requantifiée avant transmission puis de nouveau avant détection ; requantifiée uniquement avant détection. La quantication proposée renforce nettement la robustesse au bruit de la détection de l'alarme, y compris lorsqu'elle est appliquée uniquement avant la détection.

### Débruitage de parole perturbée par un bruit sporadique

Nous considérons la transmission d'un signal de parole sur un canal ajoutant un bruit sporadique survenant par salves, ce qui, malgré un RSB global élevé, peut être très gênant. La chaîne de communication, représentée sur la figure 3.18, inclut un entrelacement/désentrelacement pour disperser les erreurs à corriger.

La parole peut être considérée comme ayant une distribution laplacienne [9]. Pour faciliter les calculs, nous avons considéré un bruit également laplacien, de même variance, présent 5 % du temps. À partir de la formule (3.29), nous pouvons calculer la probabilité d'erreur pour chaque niveau de bit ; la formule (3.28) permet d'en

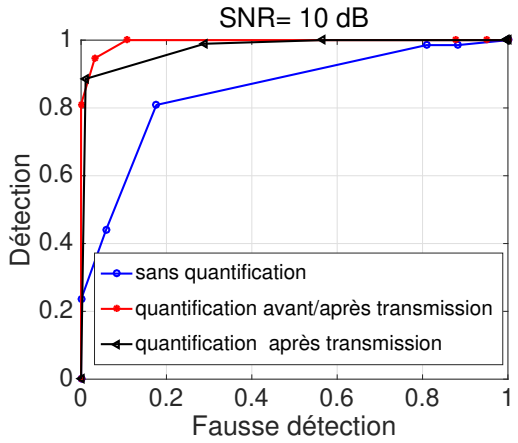



FIGURE 3.17 – Courbes ROC paramétrées par $\epsilon$, avec $\mu = 3$, pour une sirène de pompiers dans du bruit de trafic avec un rapport signal à bruit de 10 dB

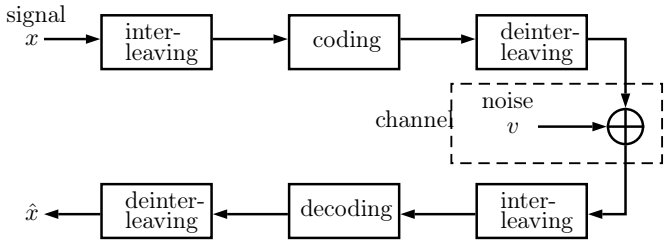



FIGURE 3.18 – Chaîne de communication *via* un canal bruité, incluant le codage proposé et un entrelacement des échantillons.

déduire la probabilité d'erreur après décodage pour chaque pouvoir de correction et chaque niveau de bit. Un exemple de ces probabilités après décodage est donnée par la figure 3.19. Ainsi, nous avons pu choisir les pouvoirs de correction des codeurs de manière à avoir une probabilité d'erreur d'environ $10^{-3}$ pour chaque poids de bit : sur l'exemple de la figure 3.19, on choisit $0, 1, 2, 3 \ldots 3$.

Les probabilités d'erreur empiriques par poids de bit avant et après décodage sont conformes aux courbes théoriques. Grâce au taux d'erreur de l'ordre de $10^{-3}$ atteint, le bruit de canal est presque supprimé, seuls quelques rares échantillons subsistant. Cependant, en considérant globalement la chaîne de communication, on a remplacé un bruit sporadique gênant par un bruit de quantification permanent. Nous avons comparé les dégradations subjectives estimées par PESQ [14] : comme indiqué par la ligne « sans reformage » du tableau 3.2, la qualité du signal codé est pauvre, mais l'introduction du processus de codage/décodage dans la chaîne de communication améliore la qualité du signal reçu.

Nous avons montré expérimentalement qu'il est possible d'améliorer cette qualité par un reformage spectral du bruit de quantification. Puisque celui-ci est blanc, il suffit de blanchir le signal à coder puis de recolorer le signal codé avec le filtre inverse, ce qui donne au bruit de quantification la même densité spectrale de puissance que celle du signal [15]. La figure 3.20 résume cette proposition. Le filtre blanchisseur devrait idéalement s'adapter au signal original, mais il doit être le même en émission et en réception, ce qui nécessiterait de transmettre ses coefficients *via* un canal auxiliaire. Pour simplifier, nous avons utilisé un filtre fixe grossièrement adapté au spectre à long terme de la parole, de fonction de transfert $1 - 0,8z^{-1}$.

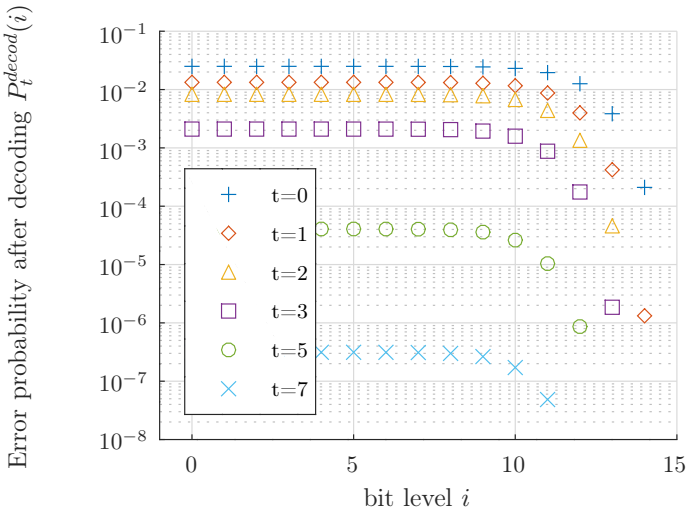



FIGURE 3.19 – Pour un signal de parole perturbé par un bruit laplacien de même variance survenant avec une probabilité de 5 %, probabilité d'erreur théorique après décodage pour chaque niveau de bit et chaque pouvoir de correction $t$, avec des codes BCH de longueur 31. La probabilité d'erreur avant décodage correspond au pouvoir de correction 0.

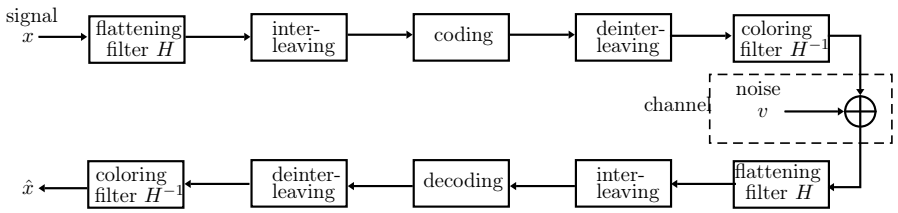



FIGURE 3.20 – Chaîne de communication avec un canal bruité, incluant le codage proposé, un entrelacement et un reformage spectral du bruit de quantification.

| | $\tilde{x}$ *vs.* $x$ | $\hat{x}$ *vs.* $\tilde{x}$ | $\hat{x}$ *vs.* $x$ | $x+v$ *vs.* $x$ |
|---|---|---|---|---|
| sans reformage | 1.98 | 2.99 | 1.83 | 1.55 |
| avec reformage | 2.51 | 2.78 | 2.14 | 1.55 |

TABLE 3.2 – Notes moyennes d'opinion (MOS) estimées par PESQ pour le signal codé $\tilde{x}$, le signal décodé $\hat{x}$ et le signal + bruit sans codage-décodage, $x+v$ .

Comme ce filtrage modifie les distributions du signal et du bruit, le modèle théorique utilisé pour choisir le pouvoir de correction pour chaque niveau de bit n'est plus valable, de sorte qu'on doit utiliser les taux d'erreur empiriques avant décodage. Les scores MOS estimés par PESQ, indiqués sur la ligne « avec reformage » du tableau 3.2, montrent une amélioration nette de la qualité du signal codé. Comme les bits de poids fort sont légèrement moins bien corrigés que dans le cas sans reformage, le gain global de qualité n'est toutefois pas aussi important, bien que meilleur que sans reformage.

Les fichiers audio peuvent être écoutés sur :
`https://helios2.mi.parisdescartes.fr/~mahe/Recherche/robustSignals/`

### 3.3.5 Pistes de travail sur la quantification auto-correctrice

Le résultat attendu de cette quantification auto-correctrice est que le signal codé-bruité-décodé soit préférable au signal bruité. La méthode proposée n'atteignant pas cet objectif en termes de rapport signal à bruit, nous l'avons atteint en considérant des signaux pour lesquels le RSB n'est pas le critère le plus pertinent : soit des signaux qui doivent simplement respecter une certaine spécification, sans référence à un signal original auquel il faudrait être fidèle, soit des signaux dont la qualité se mesure par des critères perceptifs (audio, image...). À cet égard, l'approche perceptive pourrait être exploitée non seulement dans l'évaluation, mais aussi dans la quantification elle-même, en remplaçant, dans l'algorithme 1, la maximisation du RSB (indirectement visée par la maximisation de $\langle C|R_{i-1}\rangle - \lambda 2^{L-i} w_C$) par l'optimisation d'un critère perceptif.

Que l'on utilise le RSB ou un autre critère, une autre piste de réduction du bruit de quantification réside dans la construction des dictionnaires : il s'agirait de construire des dictionnaires (pas nécessairement par poids de bits) ayant des propriétés similaires à celles des codes correcteurs d'erreur, mais adaptés aux données à coder, comme dans le codage de source. En d'autres termes, une des étapes suivantes de ce travail est de dépasser la source d'inspiration initiale qu'étaient les codes correcteurs binaires.

L'originalité et l'intérêt de ces travaux réside dans la combinaison d'approches propres au traitement du signal audio avec des techniques issues des communications numériques, deux domaines traités par des communautés scientifiques séparées. Nous avons découvert *a posteriori* une approche similaire dans les travaux d'Olivier Rioul sur la suppression du bruit « poivre et sel » dans les images [36], qui pourraient fournir des outils utiles pour la suite de ce travail.

## 3.4 Conclusion

Le nom de tatouage dopant que nous avons utilisé dans certaines publications pour caractériser l'approche présentée dans ce chapitre a pu parfois dérouter des rapporteurs experts en tatouage, tant nous nous sommes éloignés de ce que l'on considère usuellement comme tatouage audio. Dans le cas présent, on ne transmet certes plus d'information sous forme explicite de message binaire, mais on imprime une information implicite dans le signal. La contrainte d'inaudibilité de la modification du signal reste la même. Enfin, on retrouve les deux grandes familles de techniques de

tatouage : additif (reformage d'histogrammes temporels ; forçage de spectrogramme pour la SCA) et substitutif (reformage d'histogrammes temps-fréquence ; parcimonisation $\ell_0$ de spectrogramme ; quantification auto-correctrice).

Initialement prévu comme un axe de même importance que le tatouage réflexif dans le projet WaRRIS, le tatouage dopant a pris une place prépondérante dans nos travaux (bien au-delà de ce projet), liée au hasard des rencontres et à l'attrait ludique de certaines applications. Nous nous sommes progressivement rendu compte de la grande diversité du champ d'application de cette idée, même si le temps disponible pour explorer ce champ et les limites de certaines applications suggèrent un tri.

Le principal intérêt du tatouage dopant, démontré dans ce chapitre, est de rendre performants des algorithmes basiques de traitement normalement mal adaptés aux signaux audio, en reportant la complexité dans le pré-traitement que constitue le dopage. Le chapitre suivant présente une autre manière d'exploiter un tatouage pour s'adapter aux besoins d'un algorithme basique de traitement du signal.

## 3.5 Références bibliographiques

[1] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch. Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 18(1) :34–49, 2010.

[2] R. Bose and D. Ray-Chaudhuri. On a class of error correcting binary group codes. *Information and Control*, 3 :68 – 79, 1960.

[3] G. Cantisani. *Neuro-steered music source separation.* Theses, Institut Polytechnique de Paris, Dec. 2021.

[4] F. Chapeau-Blondeau and D. Rousseau. Noise-enhanced performance for an optimal bayesian estimator. *IEEE Transactions on Signal Processing*, 52(5) :1327–1334, 2004.

[5] P. Comon and P. Jutten. *Handbook of Blind Source Separation.* Academic Press, 2010.

[6] G. C. Davis and S. Mallat. Wavelet vector quantization with matching pursuit. In *Proceedings of 1994 Workshop on Information Theory and Statistics*, pages 55–, 1994.

[7] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7) :2046–2057, 2011.

[8] S. Ewert, B. Pardo, M. Muller, and M. D. Plumbley. Score-informed source separation for musical audio recordings : An overview. *IEEE Signal Processing Magazine*, 31(3) :116–124, 2014.

[9] S. Gazor and W. Zhang. Speech probability distribution. *IEEE Signal Process. Lett.*, 10(7) :204–207, July 2003.

[10] A. Gilloire and V. Turbin. Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 3681–3684, May 1998.

[11] R. Gribonval and E. Bacry. Harmonic decompositions of audio signals with matching pursuit. *IEEE Trans. on Signal Processing*, 51(1) :101–111, 2003.

[12] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3) :626–634, 1999.

[13] H. Imai and S. Hirakawa. A new multilevel coding method using error-correcting codes. *IEEE Trans. on Inf. Theory*, 23 :371–377, 1977.

[14] ITU-T. Recommendation P.862 : Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[15] N. S. Jayant and P. Noll. *Digital Coding of Waveforms, Principles and Applications to Speech and Video.* Prentice Hall, 1984.

[16] A. J. M. Kaizer. Modeling of the nonlinear response of an electrodynamic loudspeaker by a Volterra series expansion. *J. Audio Eng. Soc*, 35(6) :421–433, 1987.

[17] J. Kemp and H. Primack. Impulse response measurement of nonlinear systems : Properties of existing techniques and wide noise sequences. *J. Audio Eng. Soc*, 59(12) :953–963, 2011.

[18] W. Klippel. Modeling the nonlinearities in horn loudspeakers. *J. Audio Eng. Soc*, 44(6) :470–480, 1996.

[19] W. Klippel. Nonlinear system identification for horn loudspeakers. *J. Audio Eng. Soc*, 44(10) :811–820, 1996.

[20] W. Klippel. Loudspeaker nonlinearities- causes, parameters, symptoms. *J. Audio Eng. Soc.*, 54(10) :901–939, 2006.

[21] S. Larbi. *Structures d'égalisation en tatouage audio numérique.* Thèses, Télécom ParisTech et École Nationale d'Ingénieurs de Tunis, 2005.

[22] P. C. Loizou. *Speech Enhancement : Theory and Practice (Signal Processing and Communications).* CRC, 1st edition, June 2007.

[23] O. Macchi. *Adaptive processing : The least mean squares approach with applications in transmission.* John Wiley & Sons, 1995.

[24] V. J. Mathews. Orthogonalization of correlated Gaussian signals for Volterra system identification. *IEEE Signal Process. Lett.*, 2(10) :188–190, Oct. 1995.

[25] M. Mboup, M. Turki, and F. El Jili. Joint detection and identification of emergency vehicle alarm sound. In *12ème Colloque Africain sur la Recherche en Informatique et Mathématiques Appliquées*, pages 133–140, 10 2014.

[26] E. Z. Nadalin, R. Suyama, and R. Attux. An ICA-based method for blind source separation in sparse domains. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA2009)*, pages 597–604, Paraty, Brazil, March 2009.

[27] F. M. Naini, G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Estimating the mixing matrix in sparse component analysis (SCA) based on partial k-dimensional subspace clustering. *Neurocomputing*, 71(10-12) :2330–2343, 2008. Neurocomputing for Vision Research Advances in Blind Signal Processing.

[28] F. Nesta and M. Omologo. Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1) :246–260, 2012.

[29] B. Ninness and F. Gustafsson. A unifying construction of orthonormal bases for system identification. *IEEE Trans. Autom. Control*, 42(4) :515–521, Dec. 1997.

[30] A. Novak, L. Simon, and P. Lotton. Synchronized Swept-Sine : Theory, Application, and Implementation. *Journal of the Audio Engineering Society*, 63(10) :786–798, Nov. 2015.

[31] R. D. Nowak and B. D. V. Veen. Random and pseudorandom inputs for Volterra filter identification. *IEEE Trans. Signal Process.*, 42(8) :2124–2135, Aug. 1994.

[32] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :550–563, 2010.

[33] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4) :1118–1133, 2012.

[34] M. Parvaix, L. Girin, and J. Brossier. Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding . *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6) :1721 – 1733, Aug. 2011.

[35] J. Pinel and L. Girin. "Sparsification" of audio signals using the MDCT/IntMDCT and a psychoacoustic model – application to informed audio source separation. In *Proc. of the 42nd Audio Engineering Society Conference : Semantic Audio*, Ilmenau, Germany, 2011.

[36] O. Rioul. A spectral algorithm for removing salt and pepper from images. In *1996 IEEE Digital Signal Processing Workshop Proceedings*, pages 275–278, 1996.

[37] M. J. S. Larbi. Audio watermarking : A way to stationarize audio signals. *IEEE Trans. Signal Processing*, 53(2) :816–823, 2005.

[38] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau. Phoneme level lyrics alignment and text-informed singing voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :2382–2395, 2021.

[39] A. Stenger and W. Kellermann. Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling. *Signal Process., Elsevier*, 80(9) :1747–1760, March 2000.

[40] M. N. Syed, P. G. Georgiev, and P. M. Pardalos. A hierarchical approach for sparse source blind signal separation problem. *Computers and Operations Research*, 41(0) :386 – 398, 2014.

[41] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias. Mixing matrix estimation using discriminative clustering for blind source separation. *Digital Signal Processing*, 23(1) :9 – 18, 2013.

[42] L. Tronchin. The emulation of nonlinear time-invariant audio systems with memory by means of Volterra series. *J. Audio Eng. Soc*, 60(12) :984–996, 2012.

[43] M. Tsujikawa, T. Shiozaki, Y. Kajikawa, and Y. Nomura. Identification and elimination of second-order nonlinear distortion of loudspeaker systems using Volterra filter. In *IEEE ISCAS*, Geneva, 2000.

[44] E. Vincent and Y. Deville. *Handbook of Blind Source Separation*, chapter Audio applications, pages 779–820. Academic Press, 2010.

[45] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469, 2006.

[46] S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, 10(5) :819–829, 1992.

[47] B. Widrow. A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory. *IRE Transactions on Circuit Theory*, 3(4) :266–276, December 1956.

# Chapitre 4

# Le bruit témoin des altérations du signal

## Le tatouage témoin, pique-bœuf de l'audio

**Sommaire**



Alors que la robustesse du tatouage aux altérations du signal hôte est une contrainte forte dans de nombreuses applications du tatouage, une classe particulière d'applications repose au contraire sur la fragilité du tatouage : il s'agit des applications de vérification d'intégrité d'un document, où l'on cherche à détecter si le document a été manipulé ou falsifié. Dans ce cas, toute altération du signal doit se traduire par la modification, voire la disparition du tatouage.

Nous avons étendu ce principe en développant l'idée de *tatouage témoin* : c'est un tatouage qui s'imprègne des altérations subies par le signal hôte, de sorte que la comparaison entre le tatouage altéré et sa version originelle permet d'identifier ces altérations et, partant, de les corriger. L'objectif de *détection* assigné au tatouage fragile est ainsi étendu à une caractérisation complète des transformations du signal hôte.

Lorsque nous avons débuté cette étude, les seuls travaux (à notre connaissance) s'en approchant concernaient l'égalisation en communications numériques : en 2000, Mazzenga [8] proposait d'identifier un canal de transmission par l'insertion d'une séquence pilote cachée dans une modulation d'amplitude de porteuses en quadrature. La comparaison avec l'audio est toutefois limitée par ce que les signaux de transmissions sont soumis à des spécifications moins fortes que la contrainte d'inaudibilité du tatouage. Dans le domaine de l'audio, un pilotage de la séparation de sources par un tatouage a été proposé [5] peu après notre première publication.

Mettre en œuvre le tatouage témoin suppose de pouvoir extraire du signal tatoué distordu le tatouage distordu, pour pouvoir comparer celui-ci au tatouage original, supposé connu, et en déduire la distorsion à identifier. Cette tâche peut être facilitée si la distorsion est linéaire, si le signal hôte original est connu ou si l'on est capable de représenter les signaux dans un espace où le tatouage distordu est séparable du reste du signal. Les deux premières conditions sont vérifiées dans le cas de l'identification de système linéaire. C'est dans ce contexte, et plus précisément dans celui de l'annulation d'écho acoustique (*acoustic echo cancellation*, AEC), que nous avons étudié le tatouage dopant.

Notre réflexion est partie des travaux de Sonia Larbi et Mériem Jaïdane précédemment évoqués [13] où le tatouage d'un signal audio stationnarisait celui-ci, ce qui permettait d'améliorer les performances d'un annuleur d'écho piloté par ce signal. Pourquoi ne pas piloter l'AEC directement par le signal de tatouage, dont la blancheur et la stationnarité garantiraient les performances optimales ?

De nombreux travaux ont proposé des améliorations du schéma de base de l'AEC pour réduire la sensibilité des algorithmes adaptatifs à la non-stationnarité et à la corrélation des signaux de parole. L'introduction de filtres blanchisseurs [9] rend les algorithmes plus robustes à la corrélation. Pour compenser la non-stationnarité, les algorithmes adaptatifs à pas variable [7, 11] permettent une adaptation aux variations locales de la puissance du signal, tandis que l'algorithme de projection affine (APA [10]) réduit l'effet des variations du spectre. Comme dans les chapitres précédents, notre objectif est au contraire d'adapter le signal aux besoins d'un algorithme basique, le NLMS [6].

Dans ce chapitre, nous présentons d'abord le cheminement de l'idée initiale présentée *supra* à la structure d'AEC finalement retenue. Puis nous proposons deux réalisations de cet AEC : l'une par un algorithme adaptatif piloté par un bruit blanc gaussien stationnaire, l'autre par un corrélateur utilisant des séquences à longueur maximale. Enfin nous résumons les performances de notre AEC dans différentes conditions.

**Collaborations :**

— Mériem Jaïdane, Sonia Larbi et Monia Turki (U2S, ENIT)

— Mamadou Mboup (LIPADE, Université Paris Descartes)

**Encadrements :**

— thèse de Imen Mezghani, en co-tutelle Paris 5 - ENIT (2005-2010), co-encadrée avec Mériem Jaïdane, Sonia Larbi et Monia Turki (U2S, ENIT)

— CDD d'ingénieur d'études de Cédric Le Coz, ingénieur ENSSAT, 2009-2010

**Projets :** WaRRIS (2005-2018) et CMCU2004 (2005-2007)

**Publications :**

[DLMM+18] S. Djaziri-Larbi, G. Mahé, I. Mezghani, M. Turki, and M. Jaïdane. Watermark-driven acoustic echo cancellation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2) :367–378, Feb 2018.

[IGM+09] Mezghani-Marrakchi I., Mahé G., Jaïdane-Saïdane M., Djaziri-Larbi S., and Turki-Hadj-Allouane M. Procédé et dispositif d'annulation

d'écho acoustique par tatouage audio. brevet déposé en France le 29 octobre 2009 sous le no 0957636, étendu à l'international sous le no WO/2011/051625, 2009.

[IMS+06] Mezghani-Marrakchi I., Turki-Hadj-Allouane M., Djaziri-Larbi S., Jaïdane-Saïdane M., and Mahé G. Analyse des performances d'une nouvelle structure d'aec dans le domaine tatoué. In *Proc. of International Symposium on Image/Video Communications (ISIVC'06)*, Hammamet, Tunisia, september 2006.

[MTHADL+06] I. Marrakchi, M. Turki-Hadj Alouane, S. Djaziri-Larbi, M. Jaïdane-Saïdane, and G. Mahé. Speech processing in the watermarked domain : Application in adaptive echo cancellation. In *Proceedings of the 14th European Signal Processing Conference (Eusipco 2006)*, Florence, Italy, 2006.

[SGI+11] Djaziri-Larbi S., Mahé G., Mezghani-Marrakchi I., Turki-Hadj-Allouane M., and Jaïdane-Saïdane M. Doping and witness watermarking for audio processing. In *Proc. of the 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA 2011)*, Tipaza, Algeria, may 2011.

## 4.1 De l'idée à la structure

L'idée initiale de l'annuleur d'écho piloté par le tatouage (*watermark driven AEC*, WdAEC) est illustrée par la figure 4.1, où $F$ désigne le chemin d'écho, $w_n$ le tatouage, $x_n$ le signal de parole distant et $\nu_n$ le bruit ambiant local. Identifier le chemin d'écho $F$ à l'aide du tatouage nécessite d'éliminer, dans le signal capté $y_n$, la contribution du signal de parole $x_n$. Pour cela, nous proposons de réaliser une première identification adaptative par un AEC classique piloté par le signal tatoué $x_n^w$ (partie encadrée en pointillés). On note $G_n$ le filtre adaptatif. Dans un schéma classique d'AEC, l'erreur $e_n$ serait le signal transmis. Pour obtenir l'écho du tatouage seul (ou presque), il suffit d'ajouter à $e_n$ le tatouage $w_n$ filtré par une copie de $G_n$. On a alors :

$$e_n^w = f * w_n + \nu_n + (f - g_n) * x_n \tag{4.1}$$

Le signal $e_n^w$ contient l'écho du tatouage perturbé par le bruit ambiant et un bruit non stationnaire dont la puissance dépend de celle du signal de parole et de la convergence du premier étage d'AEC. Ce signal permet de piloter un second filtre adaptatif $G_n^w$ (ou toute autre méthode d'identification) par le tatouage seul. Les propriétés de blancheur et de stationnarité du tatouage permettent d'envisager de meilleures performances pour ce second étage, à la fois en régime transitoire et en régime permanent. On transmet alors $e_n^{tr}$, qui est la différence entre le signal capté $y_n$ et le signal tatoué filtré par une copie de $G_n^w$.

En pratique, cette structure n'offre les résultats escomptés que pour un rapport signal à tatouage (RST) insuffisant pour assurer l'inaudibilité du tatouage : pour un RST de 10 dB, la déviation quadratique moyenne est inférieure de 15dB à celle du premier étage, mais le rapport s'inverse pour un RST de 30 dB.

Nous avons donc introduit, comme illustré par la figure 4.2, un filtre $\lambda\sigma_n/A_n$ de mise en forme spectrale du tatouage avant son insertion, dont la réponse fréquentielle

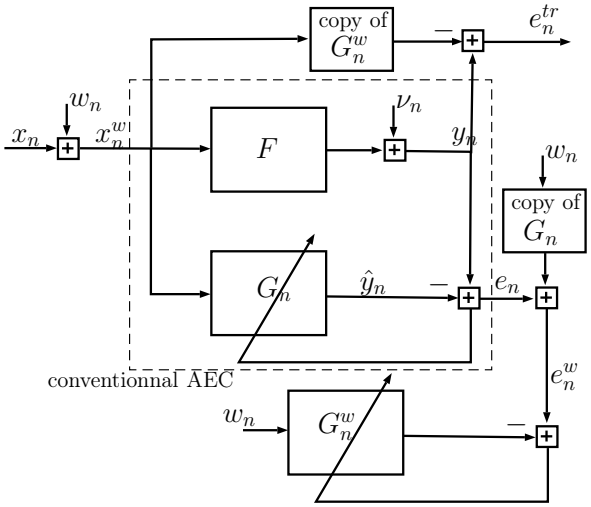



FIGURE 4.1 – Structure du WdAEC selon l'idée initiale.

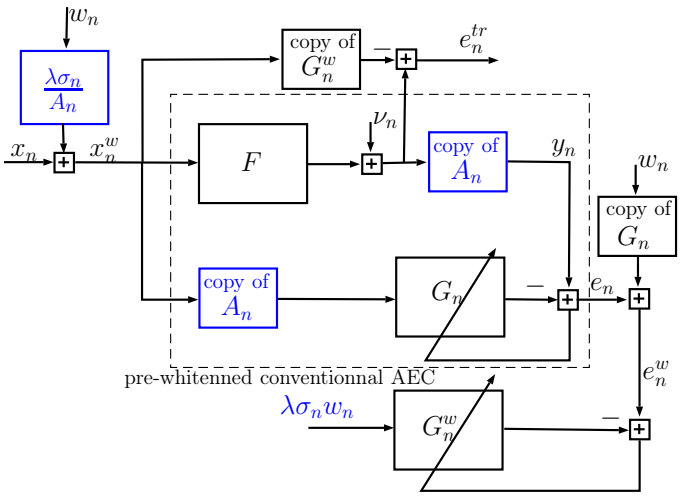



FIGURE 4.2 – Structure du WdAEC proposée dans la thèse de Imen Mezghani.

est proportionnelle à l'enveloppe spectrale du signal de parole $x_n$, tel que les coefficients de $A_n$ sont les coefficients de prédiction linéaire du signal, $\sigma_n^2$ est la variance de l'erreur de prédiction et $\lambda$ est un facteur d'atténuation garantissant l'inaudibilité du tatouage. Ce filtrage est compensé par l'introduction du filtre blanchisseur $A_n$ dans le premier étage d'AEC, de sorte que l'écho de référence du second étage, $e_n^w$, soit toujours l'écho du tatouage $w_n$ :

$$e_n^w = \lambda\sigma_n.f * w_n + a_n * \nu_n + (f - g_n) * a_n * x_n \tag{4.2}$$

Incidemment, le premier étage devient un AEC avec blanchiment du signal similaire à la structure proposée dans [9].

Dans la thèse de Imen Mezghani, nous avons montré, par des calculs théoriques complétés par des simulations, que la structure proposée assure une convergence de l'identification plus rapide que l'AEC classique piloté par le signal de parole tatoué et, en régime permanent, une déviation quadratique moyenne plus faible et plus stable et un gain d'ERLE [1] d'environ 10 dB.

1. Echo Return Loss Enhancement

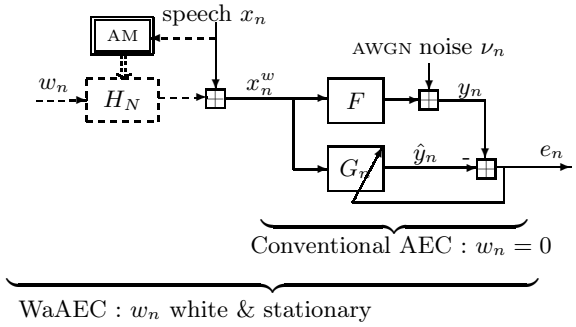



FIGURE 4.3 – AEC aidé par le tatouage (WaAEC) [4, 13], qui sera la référence et le premier étage de l'AEC proposé par la suite. AM : modèle auditif. $H_N$ : filtre perceptif de mise en forme

Cependant, en analysant plus précisément le fonctionnement de cette structure, nous nous sommes rendu compte que nous avions montré l'effet combiné du blanchiment dans le premier étage et du second étage piloté par le tatouage, sans montrer l'apport spécifique du second étage. Celui-ci s'avère faible. Avec Sonia Larbi, nous avons donc repris ces travaux et proposé une nouvelle structure, moins complexe et mettant en évidence l'intérêt du tatouage seul [DLMM+18].

Nous sommes partis de la structure proposée par Sonia Larbi dans sa thèse [4, 13], représentée sur la figure 4.3, que nous considérerons comme base de référence. Cet AEC aidé par le tatouage (*watermark aided AEC*, WaAEC) permet, grâce au blanchiment et à la stationnarisation du signal de parole par le tatouage, de respectivement accélérer la convergence de l'algorithme NLMS et améliorer de 2 à 5 dB l'ERLE en régime permanent. Le tatouage est un signal blanc et stationnaire, qui ne véhicule pas nécessairement d'information.

Nous avons proposé la structure représentée sur la figure 4.4, qui complète le WaAEC par un second étage piloté par le tatouage seul, comme expliqué ci-après. En filtrant l'erreur $e_n$ d'identification de l'écho $y_n$ par l'inverse du filtre de mise en forme, noté $H_N^{-1}$, nous obtenons :

$$e'_n = \underbrace{(f - g_n)}_{d_n} * w_n + \xi_n, \tag{4.3}$$

où :

$$\xi_n = [(f - g_n) * x_n + \nu_n] * h_N^{-1}. \tag{4.4}$$

Alors que dans les structures des figures 4.1 et 4.2 nous cherchions, dans le second étage du WdAEC, à ré-identifier le chemin d'écho $F$, nous faisons apparaître ici un nouveau problème d'identification : il s'agit d'identifier le désalignement $d_n = f - g_n$ du premier étage, le signal de référence et le bruit d'observation étant respectivement le tatouage $w_n$ et le signal $\xi_n$. On transmet alors l'écho résiduel :

$$\begin{aligned} e_n^{tr} &= e_n - \hat{d}_n * x_n^w \\ &= [d_n - \hat{d}_n] * x_n^w + \nu_n, \end{aligned} \tag{4.5}$$

où $\hat{d}_n$ est l'estimée par le 2nd étage du désalignement $d_n$ du 1er étage. Le second étage consiste soit en un filtre adaptatif, soit en une identification par bloc, que nous détaillerons respectivement dans les sections 4.2 et 4.3.

Le filtre de mise en forme $H_N$ est actualisé tous les $N$ échantillons, avec $N$ correspondant à une durée d'une vingtaine de millisecondes. Ce filtre a un gain

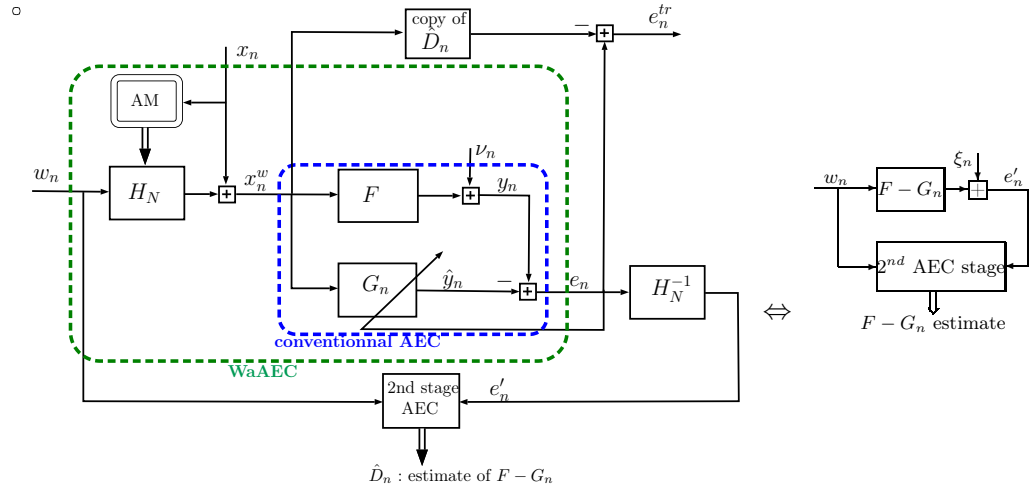



FIGURE 4.4 – Structure du WdAEC proposée dans [DLMM+18] (à gauche) et schéma équivalent d'identification pour le second étage (à droite).

d'autant plus faible que la trame de $N$ échantillons de référence est peu énergétique, ce qui accroît la puissance du bruit $\xi_n$ perturbant le second étage, réduisant les performances de celui-ci. Par conséquent, le tatouage n'est inséré que dans les trames dont l'énergie dépasse un certain seuil, ce qui a aussi un effet bénéfique sur la qualité audio du signal tatoué : en effet, la réponse fréquentielle du filtre $H_N$ n'est qu'une approximation du seuil de masquage du signal, de sorte que le tatouage de trames peu énergétiques serait audible. Dans les simulations présentées ci-après, le seuil d'énergie utilisé a conduit à un taux d'insertion de 44 % et une qualité audio du signal tatoué évaluée à un MOS de 3,5 par PESQ.

## 4.2 AEC piloté adaptativement par le tatouage (A-WdAEC)

Nous avons proposé une première implémentation du WdAEC où le second étage est un filtre adaptatif dont les coefficients sont actualisés selon l'algorithme NLMS. Le tatouage $w_n$ est un bruit blanc gaussien de variance unité.

Selon une analyse théorique classique, la vitesse de convergence dépend du conditionnement de la matrice de corrélation normalisée du signal de référence, soit $\mathbf{R_w}(n) = \mathrm{E}\left[\frac{W_n(W_n)^t}{\|W_n\|^2}\right]$ pour le second étage, contre $\mathbf{R_{x^w}}(n) = \mathrm{E}\left[\frac{X_n^w(X_n^w)^t}{\|X_n^w\|^2}\right]$ pour le premier étage (WaAEC). Comme $w_n$ est blanc alors que $x_n^w$ est fortement coloré, la vitesse de convergence du second étage doit être nettement supérieure à celle du premier.

Le comportement de l'AEC en régime permanent dépend de la puissance instantanée de $\mu\nu_n \frac{X_n^w}{\|X_n^w\|^2}$ pour le premier étage, de $\mu^w \xi_n \frac{W_n}{\|W_n\|^2}$ pour le second étage, où $\mu$ et $\mu^w$ sont les pas d'adaptation respectifs du premier et du second étages. Le bruit $\xi_n$ est plus puissant et moins stationnaire que $\nu_n$, mais cet effet est contre-balancé par ce que $\frac{W_n}{\|W_n\|^2}$ a des variations temporelles plus douces que $\frac{X_n^w}{\|X_n^w\|^2}$. Il est donc difficile de conclure théoriquement sur la supériorité de l'A-WdAEC en régime permanent.

Nous avons simulé le système dans les conditions suivantes : fréquence d'échantillonnage 16 kHz, réponse impulsionnelle de voiture de longueur 200, même longueur pour les filtres adaptatifs, RSB de 30 dB. La simulation confirme que l'A-WdAEC converge plus rapidement que le WaAEC et montre qu'en régime permanent, l'ERLE

est amélioré d'environ 10 dB. Il est remarquable que ces performances sont atteintes malgré un bruit équivalent $\xi_n$ nettement plus puissant et non-stationnaire que le bruit ambiant $\nu_n$.

## 4.3 AEC piloté par MLS (MLS-WdAEC)

La deuxième implémentation du second étage de l'AEC repose sur l'insertion dans le signal de séquences à longueur maximale (MLS), dont les propriétés de corrélation circulaire sont remarquables [14]. La réponse impulsionnelle d'un canal dont l'entrée est une MLS peut être estimée par intercorrélation entre l'entrée et la sortie. Le second étage est donc ici un corrélateur. Cette méthode a l'avantage d'être robuste au bruit et à la sous-modélisation.

Les MLS utilisées ont une longueur $L$ supérieure à la longueur $N$ des trames de parole considérées pour le reformage spectral, dont certaines ne sont pas tatouées lorsque la puissance du signal est trop faible. Ainsi, une partie des MLS sont insérées avec des poinçonnages. Nous gelons cependant l'identification du second étage lorsque moins de 20 % de la MLS est présente.

L'identification du désalignement $d_n$ du premier étage est entachée, pour chaque coefficient $d_n(l)$, de trois erreurs :

— une erreur de sous-modélisation égale à $\sum_{j=1}^{\lfloor p/L \rfloor} d_n(l+jL)$, nulle si la longueur de la réponse impulsionnelle du canal $p \leq L$ ;

— une erreur $-\frac{1}{L}\sum_{\substack{j=0 \\ j\neq l+kL}}^{p-1} d_n(j)$ due à ce que l'autocorrélation circulaire d'une MLS à un ordre différent de 0 ne vaut pas 0 mais -1 ;

— une erreur $\frac{1}{L}\sum_{k=0}^{L-1} w_k \xi_{(l+k) \bmod L}$ liée au bruit d'observation.

La deuxième et la troisième erreurs sont d'autant plus faible que $L$ est grand. Il est possible toutefois de maintenir une bonne immunité au bruit tout en réduisant $L$, en moyennant sur plusieurs $L$-périodes l'écho de référence $e'_n$ avant l'intercorrélation avec la MLS $w_n$ [12].

La simulation dans les mêmes conditions que le A-WdAEC, avec $L = 8191$, ne montre pas une convergence plus rapide que celle du WaAEC, mais l'ERLE est réduit de 5 dB, voire de 10 dB avec le moyennage de $e'$. Le choix de $L$ a un fort impact sur le troisième type d'erreur d'identification, sauf en cas de moyennage.

## 4.4 Performances comparées

La figure 4.5 compare les performances en régime permanent des différents WdAEC et du WaAEC, en considérant une modélisation exacte pour les deux étages et un bruit ambiant modéré. Comme le MLS-WdAEC identifie une réponse impulsionnelle par blocs de $L$ échantillons, les ERLE de l'A-WdAEC et du WaAEC ont été moyennés par bloc pour les besoins de la comparaison. Le meilleur ERLE est atteint par le MLS-WdAEC avec moyennage ; le MLS-WdAEC sans moyennage et le A-WdAEC ont des performances similaires, intermédiaires entre celles du WaAEC et du MLS-WdAEC avec moyennage.

Nous avons étudié par simulation la robustesse des différentes structures au bruit (RSB de 15 dB au lieu de 30) et à la sous-modélisation du premier étage (filtre à

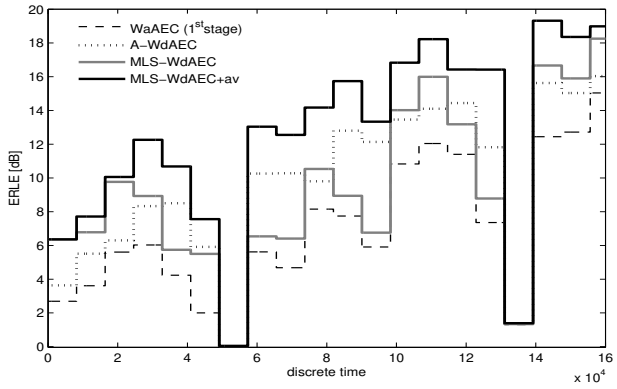



FIGURE 4.5 – Comparaison des ERLE en régime permanent des WdAEC et du WaAEC. Conditions expérimentales : réponse impulsionnelle de voiture de longueur 200, même longueur pour les filtres adaptatifs, RSB de 30 dB, pas d'adaptation 0.02 (y compris pour le 2nd étage de l'A-WdAEC), longueur des MLS $L = 8191$, 44% du signal tatoué.

100 coefficient pour une réponse impulsionnelle à identifier de 200 coefficients), dont on peut montrer l'équivalence avec un bruit. Dans les deux cas, la hiérarchie des performances est conservée, à la différence près que le MLS-WdAEC sans moyennage s'avère plus robuste que le A-WdAEC (ce qui était sa motivation initiale).

Dans l'état de l'art, les algorithmes adaptatifs d'identification sont rendus robustes au bruit en fixant leur pas d'adaptation en fonction du RSB [15, 3] ou en introduisant dans la définition du pas un terme de régularisation [2]. Nous avons volontairement écarté cette amélioration possible, l'objectif étant d'étudier l'apport du second étage de l'AEC à la robustesse au bruit, pour un pas d'adaptation du premier étage fixé et adapté à un bruit modéré. Nous avons néanmoins souhaité comparer l'efficacité de notre approche — adapter le signal aux propriétés requises par les algorithmes basiques — à l'approche de l'état de l'art — adapter les algorithmes aux « mauvaises » propriétés du signal. Pour cela, nous avons comparé le MLS-WdAEC avec un des algorithmes les plus récents selon la seconde approche, le Joint-Optimized NLMS (JO-NLMS [11]), utilisant un pas variable et une régularisation. Le choix du JO-NLMS est aussi motivé par ce que ses performances ne dépendent pas d'un réglage fin de ses paramètres, qui pourrait fausser la comparaison. La figure 4.6 illustre le gain d'ERLE du MLS-WdAEC moyenné par rapport au JO-NLMS, pour différents cas. Alors que l'ERLE du JO-NLMS est supérieur de 5 dB en modélisation exacte avec un bruit modéré, le MLS-WdAEC se révèle plus robuste au bruit et à la sous-modélisation, avec un ERLE meilleur de 5 dB dans ces cas.

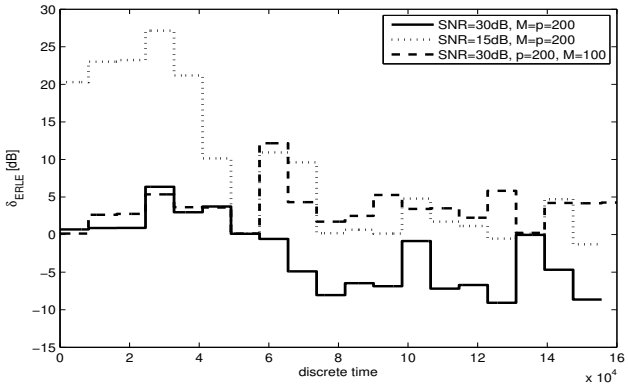



FIGURE 4.6 – Gain d'ERLE du MLS-WdAEC moyenné par rapport au JO-NLMS [11] : $\delta_{\text{ERLE}} = \text{ERLE}_{\text{MLS-WdAEC}} - \text{ERLE}_{\text{JO-NLMS}}$ dans différentes situations. $p$ et $M$ désignent les longueurs des réponses impulsionnelles respectives du canal et du filtre identificateur. Selon les notations de [11], les paramètres du JO-NLMS sont fixés comme suit : $\epsilon$=0.1, $\lambda = 1 - 1/KM$, $K = 2$, estimation de la variance du bruit selon [1].

## 4.5 Implémentation temps-réel et valorisation industrielle

Dans le cadre du projet WARRIS, nous avons recruté un jeune ingénieur de l'École Nationale Supérieure de Sciences Appliquées et de Technologies (ENSSAT) de Lannion, Cédric Le Coz, pour implémenter sous ma supervision le MLS-WdAEC sous forme d'un logiciel associé à un téléphone logiciel (*softphone*), sur un PC sous Linux. L'objectif était de disposer d'une maquette temps-réel pour faire des démonstrations, évaluer les performances subjectives par des tests formels de communications et envisager une industrialisation du WdAEC.

L'implémentation est conçue de manière à pouvoir associer l'AEC à n'importe quel softphone, de sorte que l'AEC est un programme indépendant (codé en C). Pour cela, les flux audio entre les deux applications et avec la carte son sont gérés par un serveur audio. Nous avons choisi le serveur Jack, qui avait la meilleure latence (moins de 20 ms, contre plus de 300 pour les concurrents). Cette latence dépend de la durée des trames selon lesquelles Jack découpe le signal, qui est contrainte par deux facteurs : elle doit être supérieure à la taille minimale des buffers des cartes son utilisées (32 à 64 échantillons) et au temps de traitement d'une trame. Il y a donc un compromis entre la latence et la complexité du traitement. Le choix de Jack contraignait celui du softphone : seuls Asterisk et IHearU (IHU) étaient compatibles. Asterisk ne fonctionnant qu'à 8 kHz, nous avons retenu IHU, qui ne pouvait cependant pas connecter sa sortie locale avec Jack. L'architecture matérielle et logicielle résultant des différents choix et contraintes est schématisée par la figure 4.7. Contrairement à la plupart des softphones, IHU est une application pair à pair, sans serveur, ce qui a l'avantage, dans un contexte expérimental, de faciliter le contrôle du délai de transmission.

En vue d'une valorisation industrielle du WdAEC, nous avons déposé un brevet [IGM+09]. Malgré la signature d'accords de secret avec les entreprises Proxym'IT (Tunisie), Telnet (Tunisie) et Infineon (France), le transfert technologique n'a pas abouti, par manque d'intérêt des entreprises et de de soutien institutionnel pour la valorisation.

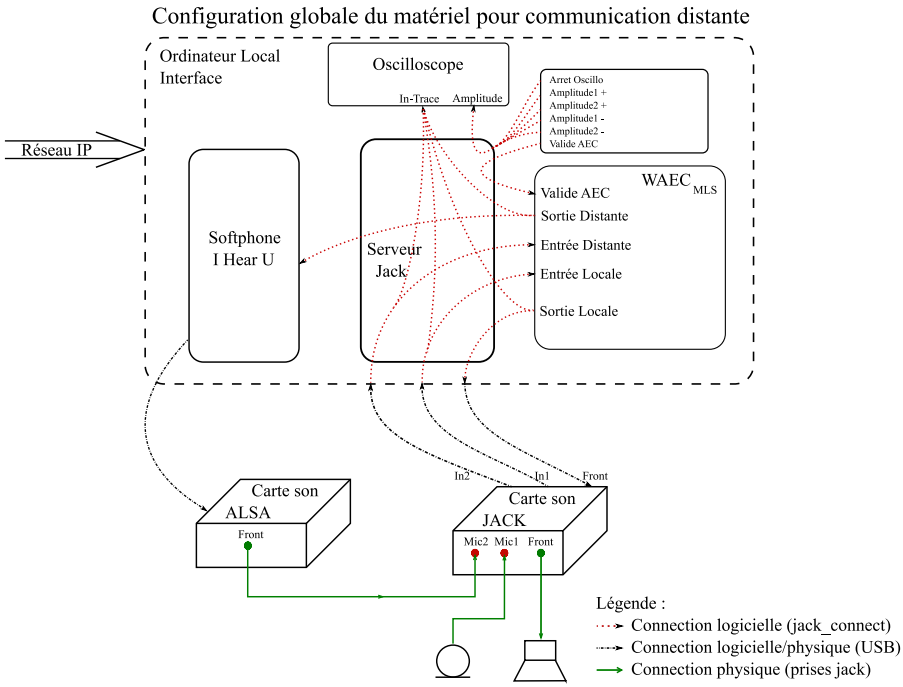



FIGURE 4.7 – Architecture matérielle et logicielle de l'implémentation temps-réel du MLS-WdAEC.

## 4.6 Conclusions

Nous avons montré qu'un tatouage doté de caractéristiques propices à l'identification de système (stationnarité et blancheur), en passant par le même canal acoustique que son signal hôte, améliore l'identification de ce canal, par comparaison entre ses versions originale et altérée.

Les performances du WdAEC sont d'autant plus remarquables que le second étage du WdAEC opère dans des conditions difficiles : le tatouage qui pilote cet étage est mis à zéro pendant la moitié du temps ; le bruit d'observation $\xi_n$ est puissant et non stationnaire ; la réponse impulsionnelle $d_n$ à identifier varie (doucement) dans le temps même si le canal acoustique est stationnaire. À cet égard, le moyennage de l'entrée du MLS-WdAEC sur plusieurs trames est contre-intuitif ; il améliore pourtant l'ERLE. Nous n'avons pas étudié la robustesse de notre système à la non-stationnarité du canal acoustique, mais cette robustesse à la non-stationnarité de $d_n$ est de bon augure.

Nous avons qualifié, en sous-titre de ce chapitre, le tatouage témoin de « *pique-bœuf de l'audio* », une analogie soulignée dans la thèse de Imen Mezghani. De même que le pique-bœuf suit le même chemin que le bœuf qui le porte, le tatouage subit les mêmes transformation que son signal hôte. De même que le pique-bœuf, par sa position, voit et analyse le chemin mieux que le bœuf tout en étant un passager de poids négligeable, le tatouage témoin a des propriétés qui lui permettent d'identifier plus facilement le canal, tout en restant inaudible.

Nous avons montré la validité du concept de tatouage témoin sur l'exemple de l'annulation d'écho acoustique, et plus largement de l'identification de systèmes linéaires convolutifs. Nous n'avons pas exploré d'autres champs d'application, car la condition d'application évoquée en introduction - pouvoir extraire le tatouage dis-

tordu du signal tatoué distordu - est difficile à vérifier. Nous avions identifié trois cas facilitants : distorsion linéaire ; signal hôte original connu ; signaux représentables dans un espace où le tatouage distordu est séparable du reste du signal. Le WdAEC s'appuie sur les deux premiers ; la suite de ces travaux pourrait consister à trouver la représentation mentionnée dans le troisième cas, sans doute liée à un contexte applicatif particulier.

## 4.7 Références bibliographiques


[1] M. Asif Iqbal and S. Grant. Novel variable step size NLMS algorithms for echo cancellation. In *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2008.

[2] J. Benesty, C. Paleologu, and S. Ciochina. On regularization in adaptive filtering. *IEEE Trans. on Audio, Speech, and Language Process.*, 19(6), 2011.

[3] H.-C. Huang and J. Lee. A new variable step-size NLMS algorithm and its performance analysis. *IEEE Trans. on Signal Process.*, 60(4), 2012.

[4] S. Larbi. *Structures d'égalisation en tatouage audio numérique.* Thèses, Télécom ParisTech et École Nationale d'Ingénieurs de Tunis, 2005.

[5] Y.-W. Liu. Sound source segregation assisted by audio watermarking. In *2007 IEEE International Conference on Multimedia and Expo*, pages 200–203, 2007.

[6] O. Macchi. *Adaptive processing : The least mean squares approach with applications in transmission.* John Wiley & Sons, 1995.

[7] V. Mathews and Z. Xie. A stochastic gradient adaptive filter with gradient adaptive step size. *IEEE Trans. Signal Process.*, 41(6), 1993.

[8] F. Mazzenga. Channel estimation and equalization for M-QAM transmission with a hidden pilot sequence. *IEEE Transactions on Broadcasting*, 46(2) :170–176, 2000.

[9] M. Mboup, M. Bonnet, and N. Bershad. LMS coupled adaptive prediction and system identification : A statistical model and transient mean analysis. *IEEE Trans. on Signal Process.*, 42(10), 1994.

[10] C. Paleologu, S. Ciochina, and J. Benesty. Variable step size NLMS algorithm for under-modeling acoustic echo cancellation. *IEEE Signal Process. Lett.*, 15, 2008.

[11] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant. An overview on optimized NLMS algorithms for acoustic echo cancellation. *EURASIP J. on Advances in Signal Process.*, (1), 2015.

[12] D. D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *J. of the Audio Eng. Soc.*, 37(6), 1989.

[13] M. J. S. Larbi. Audio watermarking : A way to stationarize audio signals. *IEEE Trans. Signal Processing*, 53(2) :816–823, 2005.

[14] D. V. Sarwate and M. B. Pursley. Crosscorrelation properties of pseudorandom and related sequences. *Proc. of the IEEE*, 68(5), 1980.

[15] L. R. Vega, H. Rey, J. Benesty, and S. Tressens. A new robust variable step-size NLMS algorithm. *IEEE Trans. on Signal Process.*, 56(5), 2008.

# Chapitre 5

# Le bruit révélateur du signal

## La NIAC

**Sommaire**



Jusqu'à présent, le bruit ajouté au signal audio pour améliorer le traitement de celui-ci restait intégré au signal, de sorte que le signal était amélioré à la fois grâce au bruit et malgré le bruit, notamment lorsque la contrainte d'inaudibilité de l'insertion n'est pas tout à fait respectée (cas du tatouage témoin ou du reformage d'histogramme, et plus encore celui de la quantification auto-correctrice). Ici, le bruit que nous ajoutons s'apparente à une poudre révélatrice d'empreintes digitales qui ne laisserait pas de trace : il s'agit d'**estimer la netteté d'un son en mesurant la sensibilité de la parcimonie du signal à un bruitage particulier**, ce bruitage étant d'autant plus inoffensif qu'il n'est effectué que dans un calcul.

**Netteté et intelligibilité**

La netteté peut être définie comme la facilité avec laquelle on peut reconnaître les différents phonèmes dans de la parole ou les notes dans une musique [38]. Elle est souvent assimilée à l'intelligibilité pour la parole, improprement selon [7] : « *La netteté concerne les mesures avec des symboles phonétiques dénués de sens ; l'intelligibilité se détermine, au contraire, avec des mots ou des phrases ayant un sens, supports ou véhicules d'une idée qu'il faut comprendre... Le premier objet de la netteté est de ne pas faire intervenir, dans la mesure, la capacité de divination que nous acquérons par la pratique de notre langue et qui nous permet de reconnaître un mot, quoique nous ne l'ayons qu'imparfaitement perçu.* »

L'état de l'art est riche et ancien en matière de prédiction de la netteté/intelligibilité par des mesures objectives, généralement spécifiques soit à la parole, soit à la musique.

Pour la parole, une première classe de méthodes est constituée des méthodes intrusives (ou avec référence) : des historiques Speech Intelligibility Index (SII [2]) et Speech Transmission Index (STI [33]) aux plus récents Short-Time Objective Intelligibility (STOI [35]) et Speech Intelligibility predictor based on Mutual Information (SIMI [21]), elles reposent sur la comparaison entre le signal distordu et sa version originale. Lorsque le signal original n'est pas disponible, on utilise des méthodes non-intrusives (sans référence). La plupart d'entre elles sont fondées sur des techniques d'apprentissage [31, 32, 1] : un indicateur est construit à partir de multiples paramètres acoustiques, en maximisant sa corrélation avec un indicateur de référence sur un corpus d'apprentissage. L'inconvénient est que ces méthodes ne reposent sur aucune analyse de ce qui fait l'intelligibilité et génèrent des indicateurs dépendant des conditions d'apprentissage. Seul le Speech to Reverberation Modulation energy Ratio (SRMR) proposé par Falk *et al.* [14] est construit différemment, en mesurant l'étalement de l'énergie de modulation vers les hautes fréquences de modulation, caractéristique de la réverbération.

La netteté de la musique, moins clairement définie, a fait l'objet de moins de travaux concernant sa prédiction. Longtemps ont dominé les mesures objectives normalisées calculées à partir des paramètres acoustiques de la pièce [20], notamment l'indice de clarté $C_{80}$, défini comme le ratio énergétique entre les 80 premières millisecondes de la réponse impulsionnelle et le reste. Récemment, des auteurs ont proposé de remplacer les énergies impliquées dans ce ratio par des grandeurs perceptives, comme les impulsions sur les nerfs auditifs [15] ou le niveau perçu [26]. Une mesure liée au signal et non à la pièce a été proposée [38], fondée sur le rapport entre les niveaux perçus respectifs des composantes directe et réverbérée d'un signal donné.

Outre la confusion netteté-intelligibilité et le traitement séparé de la parole et de la musique, la principale critique que l'on peut adresser à ces méthodes est qu'elles considèrent la netteté d'un son comme sa non-altération par le bruit, la réverbération ou une autre distorsion ; en d'autres termes elles mesurent la qualité du canal de transmission. **Il manquait 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)** [1].

1. Hossain *et al.* [17] ont proposé en 2016 (donc trop tard pour nous) un indicateur qui pourrait répondre à cet objectif, fondé sur une représentation auditive interne, qui est une sorte de norme $L^1$ du bi-spectre des spectrogrammes respectifs de l'enveloppe et de la structure temporelle fine calculées à partir du neurogramme. Cet indicateur est bien corrélé avec les scores subjectifs d'intel-

### Genèse

Mon travail sur la netteté du son est né de deux processus concomitants et indépendants. En 2011 a été élaboré le projet ICityForAll, au sein duquel je devais contribuer au paquet sur la mesure et la correction de la netteté des annonces sonores. Initialement, ma contribution devait reposer principalement sur le travail d'un.e post-doc qui aurait travaillé à partir de l'état de l'art. Parallèlement, une discussion avec Lionel Moisan sur ses mesures de netteté d'image par la *cohérence globale de phase* [5] et le *sharpness index* [4] nous a conduit à nous demander si ces outils étaient transposables à l'audio. L'incertitude du résultat n'en faisait pas de bons candidats pour le projet ICityForAll, dont le caractère industriel imposait des résultats rapides.

Les partenaires tunisiens du projet ont proposé un indicateur de netteté pertinent et adapté à la presbyacousie, le SIMforAll [28], tandis que la post-doc recrutée, Tifanie Bouchara, a ré-orienté son travail vers la mesure de saillance auditive, plus proche des compétences développées dans sa thèse. La saillance auditive, définie comme la capacité d'un son à attirer l'attention, a été envisagée comme un moyen d'améliorer l'intelligibilité des annonces sonores en les faisant « surgir » du bruit ambiant. Nous avons élaboré un protocole de validation des mesures objectives de saillance existantes, décrit dans l'encadré de la figure 5.1. L'étude sur le *sharpness index* de Moisan *et al.* transposé à l'audio menée en parallèle a finalement permis de proposer en fin de projet un indicateur de netteté du son pertinent, qui est l'objet de ce chapitre.

### De l'image au son, le bruit comme révélateur du signal

Plusieurs mesures de netteté d'image sont fondées sur l'importance de la phase d'une représentation de Fourier dans le flou perçu, le spectre de phase portant l'information sur les contours des objets [30]. La cohérence globale de phase (GPC) [5] mesure ainsi comment la régularité d'une image, définie par sa variation totale (TV), est affectée par un déphasage aléatoire. En remplaçant les images à phase aléatoire par des champs aléatoires gaussiens équivalents, Blanchet *et al.* [4, 25] ont proposé une mesure au comportement similaire mais plus simple à calculer, le Sharpness Index (SI), défini comme la sensibilité de la TV de l'image à la convolution par un bruit blanc. Ces travaux ont aussi montré que la GPC et le SI constituent des critères efficaces pour le défloutage d'image non-supervisé.

Comment transposer la GPC ou le SI aux signaux audio ? Une image nette a un gradient parcimonieux et cette parcimonie est dégradée par un déphasage aléatoire (GPC) ou par la convolution par un bruit blanc (SI), qui augmentent la TV. Au contraire, la TV d'une image floue ou bruitée est moins sensible à ces opérations. On peut observer un comportement similaire pour les signaux audio : un son net a un spectrogramme parcimonieux, composé de segments fins (schématiquement horizontaux pour les signaux harmoniques, verticaux pour les impulsions), contrairement à un son bruité ou réverbéré, dont ces segments sont étalés dans la dimension respectivement fréquentielle ou temporelle. Convoluer le signal par un bruit blanc devrait réduire fortement la parcimonie du spectrogramme d'un son pur et peu modifier celle d'un son bruité ou réverbéré. **Nous avons donc proposé comme mesure**

ligibilité. Cependant, sa complexité le rend inapproprié pour l'autre usage que nous visons, à savoir servir de critère pour des algorithmes de rehaussement du son

**de netteté la *Non-Intrusive Audio Clarity* (NIAC [2]), définie comme la sensibilité de la parcimonie du spectrogramme à la convolution du signal par un bruit blanc.** Cette définition sera précisée dans les sections 5.1 et nous validerons la NIAC en tant que mesure de netteté dans la section 5.2.

De même que le défloutage d'image fondé sur la maximisation de la la sensibilité de la TV offre des performances supérieures à celles des méthodes classiques fondées sur la minimisation de la TV [25], on peut escompter améliorer les performances d'algorithmes de traitement du son fondés sur la norme $\ell_1$ en remplaçant ce critère par la NIAC. L'étude de la NIAC comme critère de netteté sera présentée dans la section 5.3, notamment dans le cas de la séparation de sources.

**Collaborations :**

— Lionel Moisan et Mihai Mitrea (MAP5, Université Paris Cité)

— Ricardo Suyama et Giulio G.R. Suzumura (CECS, Universidade Federal do ABC)

**Encadrements :**

— stage de M2 (2012-2013, codiplômation ENIT / Paris 5) de Hejer Najjar, co-encadrée avec Lionel Moisan et Mihai Mitrea

— post-doc de Tifanie Bouchara (2012-2013), docteure de l'Université Paris Sud

**Projet :** ICityForAll

**Publications :**

[BM14] Tifanie Bouchara and Gaël Mahé. Evaluation de la saillance d'annonces vocales par un paradigme de double-tâche. In *Actes du 12ème Congrès Français d'Acoustique (CFA2014)*, pages 625–631, Poitiers, France, april 2014.

[MMM17] G. Mahé, L. Moisan, and M. Mitrea. An image-inspired audio sharpness index. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 683–687, Aug 2017.

[MSMS22] Gaël Mahé, Giulio G.R. Suzumura, Lionel Moisan, and Ricardo Suyama. A non intrusive audio clarity index (NIAC) and its application to blind source separation. *Signal Processing*, 194 :108448, 2022.

2. Nous l'avons initialement proposée sous le nom de *audio Sharpness Index* [MMM17], mais le terme de *sharpness* en audio ne désigne pas la netteté, mais l'agressivité du son

La validation subjective des modèles perceptifs instrumentaux estimant la saillance d'un signal audio reste insuffisante, car les protocoles de validation sont fondés sur des stimuli courts et artificiels. Notre objectif était de proposer un protocole de mesure subjective de la saillance auditive d'une phrase complète, permettant d'identifier les paramètres acoustiques responsables de la saillance.
Une méthode consiste à demander directement au sujet d'évaluer le niveau de saillance [22, 37, 13], ce qui pose le problème de l'interprétation de cette notion par le sujet. Une autre méthode repose sur l'idée que plus un son est saillant, plus il est détectable à un niveau faible [22, 36]. Mais cette méthode n'est adaptée qu'à des stimuli courts. Enfin, Duangudom [12] a proposé un protocole fondé sur le paradigme de double tâche [16], particulièrement pertinent dans le contexte du projet ICityForAll, qui vise les personnes presbyacousiques. En effet, les résultats des tests d'intelligibilité classiques peuvent être faussés chez ces sujets, qui compensent leur handicap par une plus grande mobilisation de leurs ressources cognitives. La baisse de performances (sur la durée) provoquée par la fatigue résultante n'est pas prise en compte. Les protocoles fondés sur le paradigme de double tâche consistent à demander au sujet d'effectuer une tâche primaire d'écoute tout en réalisant une tâche secondaire. L'effort demandé par la tâche primaire est mesuré en comparant les performances du sujet dans cette tâche avec et sans tâche secondaire. Le protocole présenté dans [12] utilise hélas des stimuli courts, non écologiques. Nous avons donc proposé un protocole fondé sur le même paradigme mais utilisant des phrases comme stimuli.
Notre protocole repose sur l'idée que pour une tâche primaire difficile, demandant de nombreuses ressources attentionnelles, augmenter la saillance du stimulus d'une tâche secondaire facile devrait augmenter le succès du sujet dans cette tâche secondaire. La tâche primaire est dérivée du test des additions en série (PASAT [11, 29]), consistant à écouter une série de chiffres séparés de 3 s et à indiquer le plus rapidement au clavier la somme de chaque chiffre avec son prédécesseur. La tâche secondaire est une tâche de discrimination parmi deux possibilités : le sujet entend toutes les 15 s une annonce de la forme : « *Le train à destination de <DESTINATION> partira à <HORAIRE> quai <LETTRE>* », où LETTRE vaut A ou B, et il doit indiquer le plus rapidement au clavier le quai, uniquement si la destination est celle qu'on lui a spécifié au début de la séquence de test.
Nos hypothèses sont les suivantes :

1. exécuter les deux tâches simultanément réduit les performances de la tâche 1 par rapport à la condition où celle-ci est exécutée seule ;
2. augmenter la saillance des annonces améliore les performances de la tâche 2, en diminuant la gêne des autres sources sonores ;
3. augmenter la saillance des annonces réduit les performances de la tâche 1 en réduisant les ressources attentionnelles consacrée à celle-ci.

Pour valider ces hypothèses et le protocole, nous avons mené une expérience pilote où nous manipulons un paramètre simple reconnu comme facteur de saillance, le niveau sonore.
L'expérience est divisée en plusieurs séries correspondant à diverses conditions : tâche 1 seule ; tâches 1 et 2 simultanément, avec les annonces à différents niveaux ; tâche 1 perturbée par les annonces sonores à différents niveaux. Cette dernière famille de conditions permet de vérifier si la diminution des performances dans la tâche 1 est due à un masquage sonore ou à une saturation des ressources cognitives.
Les résultats permettent de valider l'hypothèse 1. En revanche, même si l'augmentation du niveau sonore des annonces semble augmenter le temps de réponse dans la tâche 1 et réduire celui de la tâche 2, l'analyse statistique ne montre pas d'effet significatif.
Le protocole n'a donc pas pu être validé, mais l'analyse des résultats permet d'envisager des corrections aisées : augmenter le nombre de sujets ; augmenter les variations de niveau d'annonce d'une condition à l'autre ; réduire le délai entre deux chiffres dans la tâche primaire pour rendre celle-ci plus difficile ; réduire la proportion d'annonces correspondant à la destination-cible du sujet (la proportion de 50 % utilisée a pu en effet provoquer une contentration sur la tâche secondaire indépendante du niveau sonore des annonces).

FIGURE 5.1 – Protocole de mesure de saillance d'annonces sonores [BM14].

## 5.1 La Non-Intrusive Audio Clarity (NIAC)

### 5.1.1 Spectrogramme et parcimonie

Nous considérons l'analyse temps-fréquence d'un signal discret $s$ de durée finie $N_s$, avec des fenêtres d'analyse de durée $N$ se recouvrant de $(1-\lambda)N$ échantillons ($0 < \lambda < 1$, $\lambda N \in \mathbb{N}$). Nous définissons le spectrogramme de $s$ par :

$$S(f,t) = \sum_{n=0}^{N-1} s(t+n)h(n)C(f,n), \quad f \in \{0,1,\ldots,N_f-1\}, \quad t \in \lambda N\mathbb{Z}, \tag{5.1}$$

où la fenêtre de pondération $h$, la base de fonctions $C$ et la valeur de $N_f$ ($N$ ou $N/2$) dépend de la transformée à valeurs réelles utilisée, notée $\mathfrak{T}$ par la suite. Par exemple, pour la transformée en cosinus discrète modifiée (MDCT), $N_f = N/2$ et

$$C(f,n) = \frac{2}{\sqrt{N}} \cos\left(\frac{2\pi}{N}\Big(n+\frac{1}{2}+\frac{N}{4}\Big)\Big(f+\frac{1}{2}\Big)\right). \tag{5.2}$$

Nous mesurons la parcimonie du spectrogramme par :

$$\|S\|_1 = \sum_{f,t} |S(f,t)|. \tag{5.3}$$

La valeur $\|S\|_1$ est d'autant plus faible que le spectrogramme est parcimonieux.

### 5.1.2 Définition de la Non-Intrusive Audio Clarity

En nous inspirant du *Sharpness Index* [4, 25], nous proposons de mesurer la netteté d'un son $s$ par la sensibilité de la parcimonie de son spectrogramme à la convolution de $s$ par un bruit blanc gaussien.

Soit $s' = s * w$ et $w$ un bruit blanc gaussien centré de variance $\sigma_w^2 = 1/N_s$ (de sorte que $\|s'\|_2^2$ et $\|s\|_2^2$ aient la même espérance). Soient $S$ et $S'$ les spectrogrammes respectifs de $s$ et $s'$, le support of $S'$ étant tronqué à $N_t$, la durée de $S$.

La sensibilité évoquée *supra* peut s'exprimer *via* la probabilité $p$ que la convolution de $s$ par un bruit blanc accroisse la parcimonie :

$$p = \mathrm{Prob}\big[\|S'\|_1 \leq \|S\|_1\big]. \tag{5.4}$$

On s'attend à ce que cette probabilité $p$ soit très faible pour un signal net. En supposant que $\|S'\|_1$ soit à peu près gaussien (ce qu'on peut observer en pratique), on peut approcher $-\log p$ par :

$$-\log\left(\mathrm{Prob}\left[X \leq \|S\|_1 \;\middle|\; X \sim \mathcal{N}\big(\mathrm{E}[\|S'\|_1], \mathrm{Var}[\|S'\|_1]\big)\right]\right). \tag{5.5}$$

L'utilisation du log est motivée par ce que des valeurs telles que $p = 10^{-10000}$ ont été observées par Blanchet *et al.*, peu adaptées à la manipulation numérique.

Nous définissons cette quantité comme la *Non-Intrusive Audio Clarity* (NIAC) :

$$\mathcal{C}(s) \triangleq -\log\left(\Phi\left(\frac{\mathrm{E}[\|S'\|_1] - \|S\|_1}{\sqrt{\mathrm{Var}[\|S'\|_1]}}\right)\right), \tag{5.6}$$

où Var désigne la variance et $\Phi$ est la queue de la distribution normale :

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-x^2/2}\, dx \tag{5.7}$$

Notons que la NIAC est invariante par homothétie : $\forall\, \lambda \in \mathbb{R}, \quad \mathcal{C}(\lambda s) = \mathcal{C}(s)$. Comme cela apparaît dans l'équation (5.6), la NIAC est définie uniquement à partir du signal observé, sans référence à une version originale dont celui-ci serait la version dégradée, d'où le qualificatif « non-intrusif ».

### 5.1.3 Calcul de la NIAC

Le calcul de la NIAC telle que définie par l'équation 5.6 nécessite de calculer $\mathrm{E}[\|S'\|_1]$ et $\mathrm{Var}[\|S'\|_1]$. Nous avons montré que ces grandeurs peuvent s'exprimer aisément :

**Théorème 5.1** *L'espérance et la variance de $\|S'\|_1$ sont respectivement*

$$\mathrm{E}[\|S'\|_1] = \sqrt{\frac{2}{\pi}} N_t \sum_{f=0}^{N_f-1} \sigma_{S'}(f) \tag{5.8}$$

$$\mathrm{Var}[\|S'\|_1] = \frac{2}{\pi} \sum_{\substack{0 \le f, f' \le N_f - 1 \\ 1 - N_t \le \Delta \le N_t - 1}} (N_t - |\Delta|) \sigma_{S'}(f) \sigma_{S'}(f') \omega\left( \frac{\Gamma_{S'}(f, f', \Delta\lambda N)}{\sigma_{S'}(f)\sigma_{S'}(f')} \right), \tag{5.9}$$

*où*

- *$N_t$ et $N_f$ sont les nombres respectifs de colonnes et de lignes de $S$ ;*
- *$\Gamma_{S'}(f, f', \tau) \triangleq \sigma_w^2 \mathfrak{T}[\tilde{R}_{s,\tau}(n, n')]$ ;*
- *$\tilde{R}_{s,\tau}(n, n') \triangleq R_s(\tau + n - n')h(n)h(n')$, où $R_s$ désgine l'auto-correlation de $s$ (finie et déterministe) ;*
- *$\sigma^2_{S'}(f) \triangleq \Gamma_{S'}(f, f, 0)$ ;*
- *$\forall\, x \in [-1, 1], \quad \omega(x) \triangleq x \arcsin x + \sqrt{1 - x^2} - 1$.*

Le calcul de la NIAC a une complexité de $\Theta(N_t N_f^2 \log_2 N_f)$ multiplications. Avec une machine multi-processeurs, les calculs peuvent être parallélisés sur les $2N_t$ valeurs de $\Delta$, ce qui réduit la complexité temporelle à $\Theta(N_f^2 \log_2 N_f)$.

Nous avons également montré que la NIAC d'un mélange linéaire non-convolutif, qui sera utile dans la section 5.3, peut s'exprimer simplement à partir de quantités similaires :

**Théorème 5.2** *Soit $y$ une combinaison linéaire de $p$ signaux $x_1 \dots x_p$ :*

$$y = \sum_{i=1}^{p} \alpha_i x_i. \tag{5.10}$$

*La NIAC de $y$ peut être calculée en utilisant l'équation (5.6) et le théorème 5.1, avec*

$$\Gamma_{Y'}(f, f', \tau) = \sum_{1 \le i, j \le p} \alpha_i \alpha_j \Gamma_{X'_i X'_j}(f, f', \tau), \tag{5.11}$$

*où*

— $\Gamma_{X_i'X_j'}(f, f', \tau) \triangleq \sigma_w^2 \mathfrak{T}[\tilde{R}_{x_i x_j, \tau}(n, n')]$ ;

— $\tilde{R}_{x_i x_j, \tau}(n, n') \triangleq R_{x_i x_j}(\tau + n - n')h(n)h(n')$, *où* $R_{x_i x_j}$ *désigne l'intercorrélation entre* $x_i$ *et* $x_j$ *(finie et déterministe).*

En considérant ici les équivalents asymptotiques, la complexité de la NIAC du mélange connaissant les $(\Gamma_{X_i'X_j'})_{i,j}$ est de $O(p^2 N_t N_f^2)$. S'il est possible de paralléliser les calculs, la complexité est réduite à $O(4p^2 + 2)N_f^2$.

### 5.1.4 Paramétrage

La NIAC a l'avantage de dépendre de peu de réglage de paramètres : seuls ceux du spectrogramme et la durée d'analyse $T$ doivent être fixés.

Dans le calcul du spectrogramme, la longueur des fenêtres d'analyse fréquentielle est fixée classiquement autour de 20 ms, pour assurer un bon compromis entre les résolutions temporelle et fréquentielle. Utiliser des fenêtres plus longues rendrait le spectrogramme moins sensible à l'étalement dans la dimension temporelle en cas de réverbération, tandis que des fenêtres plus courtes réduiraient la résolution fréquentielle, notamment pour des signaux harmoniques, ce qui réduirait la sensibilité au bruit.

Le choix de la durée $T$ du spectrogramme est guidé par le critère de stabilité de la NIAC au cours du temps : dans des conditions constantes de production (élocution ou jeu musical) et de canal acoustique (bruit, réverbération, ...), la NIAC devrait peu varier, si elle mesure la netteté. La valeur pertinente de $T$ dépend du rythme du signal : la convolution par un bruit blanc modifie plus fortement le spectrogramme si une non-stationnarité apparaît pendant cette durée $T$, ce qui produit une NIAC élevée. Si $T$ est inférieure à la période correspondant au rythme (nombre moyen de syllabes ou de notes par seconde), certains $T$-blocs contiendront un changement, d'autres non, de sorte que la NIAC aura une moyenne faible et une forte variance. Au contraire, des valeurs élevées de $T$ conduisent à une NIAC élevée et stable au cours du temps, puisque chaque $T$-bloc est susceptible de contenir un changement. C'est ce que nous avons observé expérimentalement. Si l'on considère le ratio écart-type sur moyenne de la NIAC en fonction de $T$, on observe une brusque décroissance autour d'une valeur de $T$ correspondant au rythme du signal. Pour la parole, $T$ doit donc être choisi supérieur à la durée moyenne d'une syllabe (environ 400 ms en français [24]).

Toutefois, il est possible de moyenner la NIAC sur plusieurs blocs pour compenser la variabilité : ainsi, un moyennage sur 2 s conduit à la même stabilité au cours du temps pour toutes les valeurs de $T \leq 2$s, ce qui permet d'adapter $T$ aux autres contraintes (complexité ou utilisation de la NIAC à des fins de correction du signal au fil de l'eau).

## 5.2 La NIAC comme mesure de netteté

### 5.2.1 Comparaison avec le STI

Nous avons comparé la NIAC avec le Speech Transmission Index (STI [33]), qui est une mesure intrusive utilisée depuis longtemps et bien corrélée avec la netteté

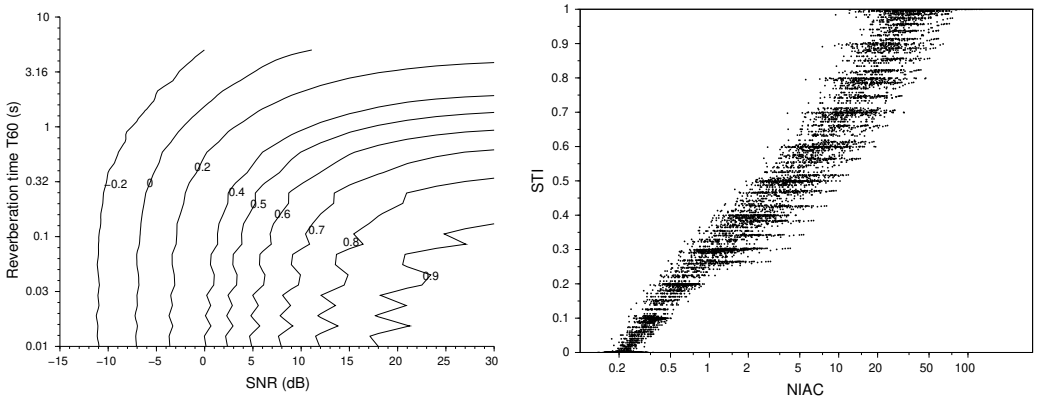



FIGURE 5.2 – (a) Lignes iso-log(NIAC) dans le plan SNR-T60 : pour chaque condition (SNR,T60), la NIAC est moyennée sur 16 locuteurs. (b) Relation entre la NIAC et le STI : chaque point représente une condition (locuteur, SNR, T60), avec 16 locuteurs, T60 prend 30 valeurs logarithmiquement réparties entre 10 ms and 5 s, et SNR prend 21 valeurs linéairement réparties entre -30 et +30 dB.

perçue. Pour cela, nous avons dégradé les signaux d'un corpus de parole et d'un corpus de musique par du bruit blanc et de la réverbération, avec différents rapports signal à bruit (RSB) et différents temps de réverbération (T60). Pour chaque extrait sonore et chaque condition (RSB, T60), nous avons calculé d'une part la NIAC, d'autre part le STI selon [18]. Bien que le STI soit initialement prévu pour mesurer l'intelligibilité de la parole, son principe (mesurer comment le canal acoustique réduit l'indice de modulation pour différentes fréquences de modulation dans différentes bandes de fréquences) en fait une mesure adaptée à n'importe quel signal audio, dès lors que les fréquences de modulation sont entre 0,63 et 12,5 Hz, ce qui est le cas de la musique.

Pour le corpus de parole, les lignes iso-log(NIAC) dans le plan (RSB,T60) (figure 5.2.a) sont similaires aux lignes iso-STI présentées dans [18]. Comme illustré par la figure 5.2.b, le log de la NIAC est en effet très corrélé au STI (coefficient de corrélation de 0,99).

Pour la musique, on peut aussi observer cette corrélation, mais la corrélation et la relation linéaire entre les deux mesures dépendent de l'instrument, de l'instant de mesure, de la durée $T$ du spectrogramme et de la durée de moyennage. En choisissant $T = 256$ ms et un moyennage sur 4096 ms, la corrélation est élevée pour tous les instruments (supérieure à 0,8) et peu dépendante de l'instant de mesure choisi.

La NIAC peut donc prédire le STI avec une grande fiabilité dans ces conditions bruit+réverbération, tout en ayant l'avantage d'être non-intrusive.

### 5.2.2 Validation subjective formelle

Dans le cadre du projet ICityForAll, une campagne de tests subjectifs a été menée conjointement par le CEA-Linklab (Tunis) et le laboratoire de traitement du signal (LTS2) de l'EPFL, en 2015. Ces tests visaient à la fois à évaluer un algorithme de pré-compensation du son (renforcement de la netteté avant diffusion) développé dans le projet et à valider les deux mesures instrumentales de netteté du son, la NIAC et le SIMforAll [28] développé au Linklab. Le SIMforAll est une mesure intrusive dérivée

du Speech Transmission Index (STI), qui y intègre une approche pychoacoustique, notamment la prise en compte de la presbyacousie.

Le test utilise le corpus du *Hearing Test In Noise* (HINT), composé de 5 listes de 20 phrases phonétiquement équilibrées, de 1,5 à 2,5 s chacune, regroupées par deux dans notre test. Toutes les phrases sont prononcées par le même locuteur masculin. Les fichiers audio des trois premières listes ont été traités par l'algorithme de pré-compensation avec trois niveaux différents respectifs de pré-compensation. Les 10 fichiers de chacune des 5 listes ont ensuite subi un bruitage par du *babble noise* (bruit de cafétéria) et une réverbération, à différents niveaux.

Vingt sujets - 10 normo-entendants et 10 malentendants - ont passé le test, consistant à écouter les double-phrases au casque à 75 dB SPL et à répéter ce qu'ils et elles avaient compris. Pour chaque double-phrase, le score d'intelligibilité est défini comme le pourcentage de mots reconnus.

La figure 5.3 présente les corrélations entre le score subjectif et les mesures objectives : SIMforAll, STI et NIAC. Le SIMforAll apparaît nettement comme le plus corrélé aux mesures subjectives d'intelligibilité. Ces résultats doivent cependant être nuancés :

— Ce test test d'intelligibilité est peu adapté à la NIAC, conçue comme une mesure de netteté, comme d'ailleurs le STI. Houtgast *et al.* [18] comparent le STI non pas à une mesure d'intelligibilité mais à une mesure de netteté, le pourcentage de non-reconnaissance de consonnes (ALcons) dans un test consonne-voyelle-consonne. Ce test est nettement plus discriminant que le test HINT pour la reconnaissance fine, lorsqu'on reconnaît plus de 80 % des mots. C'est pourquoi les scores ALcons sont présentés en échelle log (pour zoomer sur les 20 % restant). Alors que le présent test fait apparaître des STI inférieurs à 0,5 et sans corrélation avec le taux de reconnaissance lorsque celui-ci dépasse 80 %, [18] montre des STI entre 0 et 1 très bien corrélés avec l'ALcons en échelle log.

— Le bruit utilisé pour le test a un effet non négligeable. Quand le RSB est négatif, la NIAC et le STI mesurés sont ceux du bruit. Or ce bruit a des caractéristiques proches de celles de la parole, ce qui peut tromper ces deux indicateurs, fondés sur des propriétés de la parole.

En conclusion, ce test a été conçu de manière adaptée au projet : il s'agit pour des personnes presbyacousiques de reconnaître la plus grande proportion possible d'annonces dans un environnement réverbérant avec du bruit de conversations à un niveau élevé, avec un pré-traitement éventuel de ces annonces. Il n'était cependant pas adapté à l'évaluation subjective de la NIAC à ce stade de son développement. Une première évaluation subjective de la NIAC nécessiterait (i) de ne pas utiliser d'emblée un type de bruit unique dont on sait qu'il la mettra en difficulté ; (ii) de réaliser des tests de netteté ; (iii) d'utiliser plusieurs voix avec des différences marquées de clarté d'élocution, de manière à mesurer la netteté intrinsèque, et pas seulement la dégradation liée au bruit et à la réverbération.

Les résultats de ce test faisant toutefois peser quelques doutes sur la validité de la NIAC comme mesure de netteté, nous avons alors concentré nos efforts sur la validation de la NIAC comme critère de netteté pour piloter des applications de rehaussement du son.

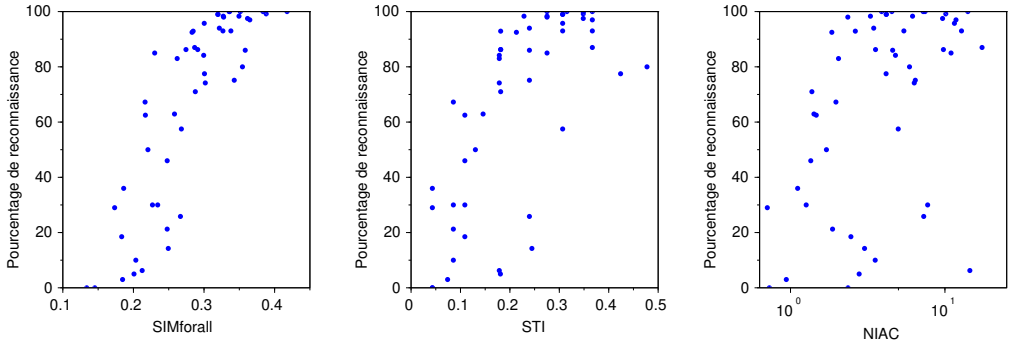



FIGURE 5.3 – Nuages de corrélation entre le pourcentage de mots reconnus et respectivement, le SIMforAll, le STI et la NIAC. Chaque point représente une double phrase dans une condition (pré-compensation, RSB, T60), avec RSB entre -8 et 4 dB et T60 de 0,5, 2 ou 4 s. Le pourcentage est moyenné sur les 10 sujets normo-entendants.

## 5.3 La NIAC comme critère de netteté

La NIAC s'exprimant par une formule analytique, elle se prête à des applications de rehaussement du son (débruitage, déréverbération, annulation d'écho, séparation de sources, *impainting*...) par optimisation de la NIAC, en considérant que le son corrigé est plus net, sous l'hypothèse que la NIAC mesure la netteté.

Nous nous sommes d'abord inspiré du cas simple traité dans [25], la déconvolution paramétrique d'un signal convolué par un noyau gaussien et bruité. Dans ce cas, il s'agit de trouver les valeurs des deux paramètres du filtre de déconvolution — rayon de floutage et facteur de régularisation — qui maximisent la NIAC. Les résultats obtenus sont très satisfaisants, mais si cette déformation modélise de manière réaliste le floutage d'une image, elle ne correspond à aucun scénario réaliste en audio. Nous avons alors cherché à corriger l'équivalent audio du floutage d'image : une réverbération pure, dont la réponse impulsionnelle est modélisable comme un bruit blanc multiplié par une exponentielle décroissante. La déconvolution est ici non-paramétrique et la réponse fréquentielle du filtre de déconvolution est obtenue par maximisation de la NIAC. Plusieurs obstacles sont apparus, qui nous ont amené à laisser provisoirement de côté ce problème, sur lequel nous revenons actuellement, après avoir traité celui de la séparation de sources.

### 5.3.1 Séparation de sources par optimisation de la NIAC

Nous considérons un mélange instantané linéaire et déterminé de $p$ signaux. En notant $s$ le vecteur des $p$ sources et $x$ le vecteur des $p$ mélanges, le mélange peut s'écrire $x = As$, où $A$ est une matrice non-singulière $p \times p$. Le but de la séparation de sources non-supervisée (BSS) est d'estimer $s$ sans connaître $A$.

Notre idée initiale était qu'une source seule est plus nette que mélangée à d'autres, de sorte que la séparation pourrait être pilotée par la maximisation de la NIAC, sous l'hypothèse que celle-ci mesure la netteté. Cependant, nos premières expériences ont montré que cette idée n'est correcte que si tous les signaux du mélange ont des NIAC du même ordre de grandeur. Dans le cas contraire, un signal à faible NIAC voit celle-ci augmenter s'il est perturbé par l'ajout d'un signal ayant une NIAC beaucoup plus forte, de sorte que son extraction correspond à la minimisation de la NIAC. Extraire

une des sources du mélange revient donc à trouver

$$\hat{\alpha} \in \arg\max_{\alpha} \overline{\mathcal{C}}(y_\alpha) \cup \arg\min_{\alpha} \overline{\mathcal{C}}(y_\alpha), \text{ avec } y_\alpha = \sum_{i=1}^{p} \alpha_i x_i,\ \alpha = [\alpha_1 \dots \alpha_p]^\top \quad (5.12)$$

où $\overline{\mathcal{C}}$ désigne la moyenne de $\mathcal{C}$ sur plusieurs blocs de durée $T$.

Comme la NIAC est invariante par homothétie, les solutions de l'équation (5.12) sont définies à un facteur d'échelle près. Nous supprimons ce degré de liberté en imposant $\mathrm{E}[y_\alpha^2] = 1$. D'autre part, les contributions de la source estimée $y_\alpha$ à chaque composante de $x$ doivent avoir le même signe, car une inversion de signe signifierait un déphasage, donc un mélange non-instantané. Soit $\hat{a}$ le vecteur des contributions estimées :

$$\hat{a} = \arg\min_{a} \mathrm{E}[\|x - y_\alpha a\|^2] = \mathrm{E}[y_\alpha x]/\mathrm{E}[y_\alpha^2]. \quad (5.13)$$

Les contraintes de signe s'expriment :

$$\pm\mathrm{E}[y_\alpha x] \ \geq\ 0, \quad (5.14)$$

Finalement, en notant $C_x = \mathrm{E}[xx^\top]$ la matrice de corrélation de $x$ et en définissant $\sqrt{C_x}$ telle que $C_x = \sqrt{C_x}\sqrt{C_x}^\top$, le problème d'optimisation à résoudre est le suivant :

$$\hat{\beta} \in \arg\max_{\beta} \overline{\mathcal{C}}(y_\beta) \cup \arg\min_{\beta} \overline{\mathcal{C}}(y_\beta) \quad \text{avec} \quad y_\beta = x^\top \left(\sqrt{C_x}^\top\right)^{-1} \beta \quad (5.15)$$

$$\text{sous les contraintes } \begin{cases} \|\beta\| = 1 \\ \text{et } \pm\sqrt{C_x}\beta \ \geq\ 0. \end{cases} \quad (5.16)$$

On doit donc optimiser une fonction sur une région d'une sphère de rayon 1 définie par des inégalités linéaires. Comme chaque point $\beta$ de la sphère est équivalent à son symétrique $-\beta$, une seule hémisphère doit être explorée. Notons que pour chaque évaluation de $\overline{\mathcal{C}}(y_\beta)$ au cours du processus d'optimisation, le théorème 5.2 permet de limiter les calculs, tous les $\Gamma_{X'_i X'_j}$ étant calculés une fois au début.

Comme nous savons que l'optimum escompté respecte la contrainte de signe $\pm\sqrt{C_x}\beta \ \geq\ 0$, on peut vérifier celle-ci *a posteriori*, après convergence de l'algorithme d'optimisation. D'autre part, la contrainte $\|\beta\| = 1$ peut être satisfaite en laissant $\|\beta\|$ libre pendant l'optimisation et en normalisant la solution à la fin ou quand on a besoin de contrôler l'algorithme d'optimisation (voir sous-section 5.3.2).

### 5.3.2 Processus d'optimisation et de séparation

Une idée que nous avons explorée est de chercher en parallèle tous les extrema de la fonction à optimiser, grâce à la *Multi-Optima Particle Swarm Optimization* (MOPSO) [8]. Outre son coût en calculs, un inconvénient de cette solution est que tous les extrema ne correspondent pas nécessairement à l'extraction d'une source. On le voit notamment dans l'exemple voix+voix de [MMM17], où chaque voix correspond à un maximum, de sorte que les deux minima locaux ne correspondent à rien.

Nous avons privilégié un processus itératif d'extraction-déflation [10]. À chaque itération, on extrait une source par maximisation ou minimisation de $\overline{\mathcal{C}}(y_\beta)$, puis on estime sa contribution au mélange pour la soustraire, et finalement on réduit la

dimension du mélange : voir l'algorithme 2. Les réductions successives de dimension réduisent d'autant plus la complexité que celle de la NIAC d'un mélange est une fonction quadratique du nombre de sources (voir 5.1.3).

Une mauvaise extraction à une itération peut toutefois compromettre les extractions suivantes. Pour limiter ce risque, nous exploitons la possibilité de maximiser ou minimiser la NIAC. Ainsi, à chaque itération, nous gardons le sens d'optimisation de la précédente (maximisation ou minimisation) et nous mesurons l'indépendance entre la source extraite et le signal résiduel. Si celle-ci n'est pas suffisante ou que la contrainte de signe (5.16) n'est pas respectée, nous testons l'optimisation dans l'autre sens et nous conservons la solution respectant la contrainte de signe et maximisant l'indépendance.

**Algorithme 2** : Processus itératif d'extraction-déflation.

$\tilde{x} \leftarrow x$
**repeat**
  Extract $y_{max}$ through NIAC maximization
  Estimate the contribution $\hat{a}^{max}$ of $y_{max}$ to $\tilde{x}$ (see Eq. (5.13))
  Deflation : $\tilde{x} \leftarrow \tilde{x} - \hat{a}^{max} y_{max}$
  Dimension reduction : write

$$\tilde{A} = \left( \begin{array}{c|c} \hat{a}^{max} & \begin{matrix} I_{\tilde{p}-1} \\ 0_{1,\tilde{p}-1} \end{matrix} \end{array} \right), \text{ where } \tilde{p} = \dim(\tilde{x})$$

  Decompose $\tilde{A}$ under the form $\tilde{A} = QR$, with $Q$ orthogonal and $R$ upper triangular
  $\tilde{Q} \leftarrow Q$ without its first column
  $\tilde{x} \leftarrow \tilde{Q}^{\top} \tilde{x}$
**until** $\dim(\tilde{x}) = 1$

Une méthode classique d'optimisation continue, comme la méthode de Newton, converge vers l'optimum rapidement et précisément si le gradient et le Hessien de la fonction à optimiser peuvent être calculés ou estimés. Cependant, elle est sensible au point d'initialisation et peut être « piégée » dans un optimum local. À l'inverse, l'optimisation par essaim de particules (*Particle Swarm Optimisation*, PSO [23]) permet de trouver l'optimum global par sa capacité à explorer tout l'espace mais converge lentement vers la position précise de l'optimum. C'est pourquoi nous avons adopté un schéma d'optimisation en deux étapes : nous cherchons grossièrement l'optimum par PSO, puis utilisons cette solution comme initialisation d'un algorithme de type Newton, ce qui accélère la convergence de celui-ci tout en limitant le risque de trouver un optimum local.

L'application des algorithmes (MO)PSO a été étudiée par Giulo Suzumura, doctorant sous la direction de Ricardo Suyama. Le principe est de répartir dans l'espace de recherche une collection de particules représentant des solutions possibles et, par une règle d'adaptation de la position et de la vitesse de chaque particule, de faire converger l'essaim de particules vers l'optimum pour PSO (ou des essaims vers les optima pour MOPSO). Le vecteur vitesse de chaque particule est modifié à chaque itération en fonction de la meilleure position trouvée jusqu'à présent par la particule et de la meilleure position trouvée par l'essaim. Nous arrêtons l'algorithme lorsque l'essaim a atteint une inertie [3] suffisante.

3. distance quadratique moyenne entre les particules et le barycentre de l'essaim

Pour la seconde étape d'optimisation, comme la fonction $t \mapsto -\log \Phi(t)$ est croissante, l'optimisation de $\mathcal{C}(y_\alpha)$ peut être avantageusement remplacée par celle de l'opérande de $\Phi$ dans l'équation (5.6), que nous appelons *pseudo-NIAC* :

$$p\mathcal{C}(s) \triangleq \frac{\mathrm{E}[\|S'\|_1] - \|S\|_1}{\sqrt{\mathrm{Var}[\|S'\|_1]}} \tag{5.17}$$

À cause de la norme $L^1$ du spectrogramme, le gradient de cette fonction est indéfini pour tout $\alpha$ tel qu'il existe un couple temps-fréquence $(t, f)$ pour lequel $\sum_{j=1}^{p} \alpha_j X_j(f,t) = 0$. La discontinuité est cependant noyée dans une somme sur tous les indices temporels et fréquentiels du spectrogramme, de sorte ces points exceptionnels n'ont pas perturbé la convergence de l'algorithme en pratique.

Pour éviter le calcul du Hessien de $p\mathcal{C}$, nous avons utilisé l'algorithme de quasi-Newton avec l'approche BFGS (QN-BFGS). Le seul paramètre à fixer est le seuil $\varepsilon$ tel que la condition $\|\beta^{(k)} - \beta^{(k-1)}\| < \varepsilon$ provoque l'arrêt de l'algorithme (avec $\beta^{(j)}$ la valeur de $\beta$ à l'itération $j$). Nous avons montré que ce seuil peut être aisément fixé selon la qualité de la séparation souhaitée, définie par le ratio signal à erreur.

### 5.3.3 Résultats expérimentaux et discussion

#### Un exemple à 3 sources

Soit un mélange de trois sources : guitare acoustique, voix et piano, avec pour matrice de mélange

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

La géographie du problème d'optimisation est illustrée par la figure 5.4.

Nous le séparons par la méthode décrite *supra*. La durée des spectrogrammes est fixée à 256 ms et la NIAC est moyennée sur 4096 ms, conformément aux résultats de la sous-section 5.2.1. La PSO utilise 10 particules. Dans l'algorithme quasi-Newton, le critère d'arrêt $\varepsilon$ est fixé à $10^{-4}$, ce qui correspond à un ratio signal à erreur (SER) cible de 80 dB.

La PSO approche le maximum en 4 itérations et le résultat sert d'initialisation à l'optimisation QN-BFGS, qui converge en 10 itérations (16 appels), avec un SER de 49 dB : voir figure 5.5. Après extraction de la voix et déflation, la PSO converge vers le maximum en 3 itérations, puis l'algorithme QN-BFGS initialisé par le résultat de la PSO converge en 4 itérations (10 appels), avec un SER de 28 dB : voir figure 5.6. On extrait alors la guitare et il reste le piano.

Nous avons évalué les performances de la séparation par les mesures suivantes : signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) et signal-to-artifact ratio (SAR) [40], présentées dans la table 5.1. Notons que les sources auraient aussi pu être séparées par minimisation de la NIAC (dans les deux itérations), conduisant à l'extraction du piano en premier. Cependant, la figure 5.6 indique une légère différence entre le minimum et le point d'extraction du piano, ce que produirait des erreurs plus importantes.

La visualisation dynamique du processus d'optimisation et les fichiers audio sont disponibles sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/HDR/`.

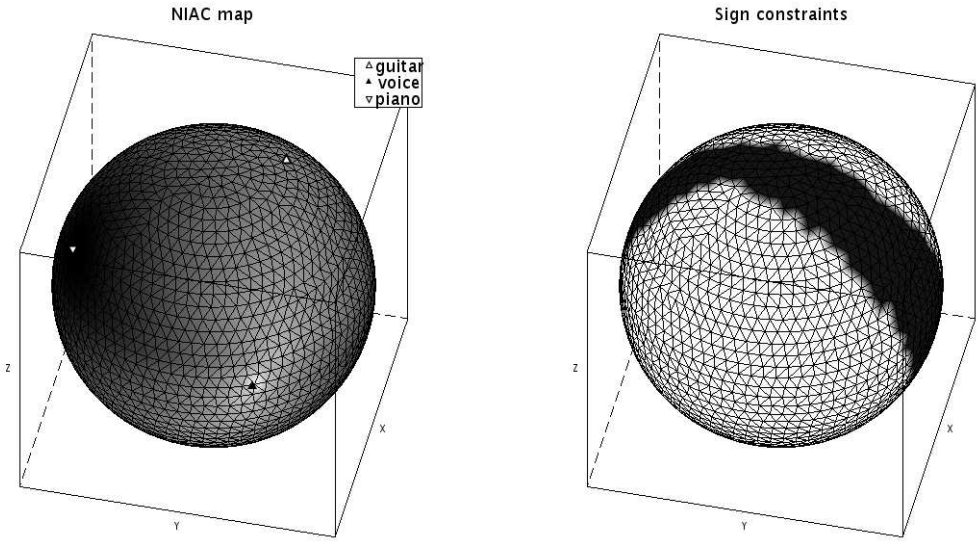



FIGURE 5.4 – Pour un mélange guitare acoustique, voix et piano : à gauche, pseudo-NIAC de $y_\beta$ (voir equation 5.15) sur la sphère de rayon 1 dans l'espace de $\beta$ et points d'extraction. ; à droite, zone interdite par les contraintes de signe, en noir.

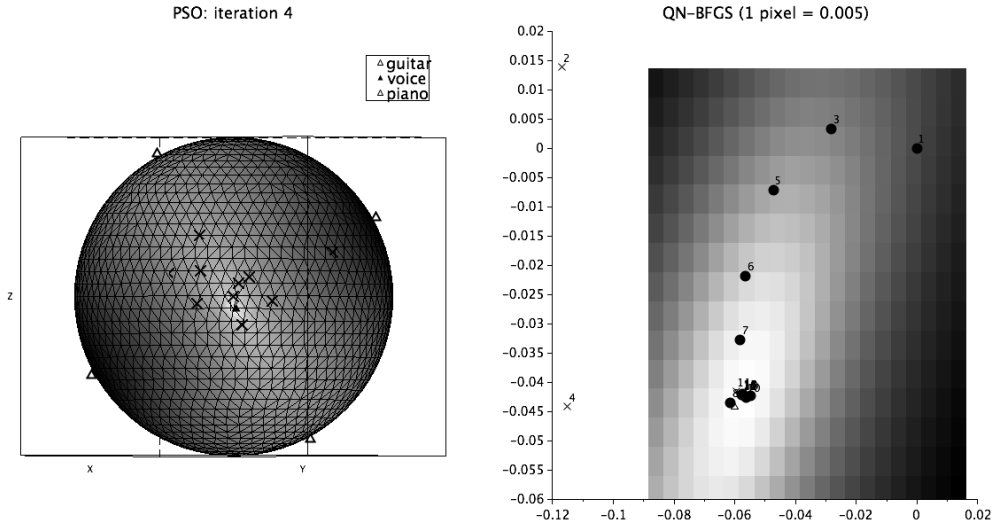



FIGURE 5.5 – Processus d'optimisation pour un mélange de 3 sources. À gauche, NIAC sur la sphère de rayon 1 dans l'espace de $\beta$, points d'extraction et essaim de particules à la dernière itération de la PSO. Les particules sont rassemblées autour du point maximisant la NIAC, correspondant à l'extraction de la voix. À droite, zoom autour de ce maximum, montrant les points successifs de l'optimisation QN-BFGS. Les points intermédiaires de la recherche linéaire non retenus sont représentés par des croix, certains sont hors de la zone de zoom.

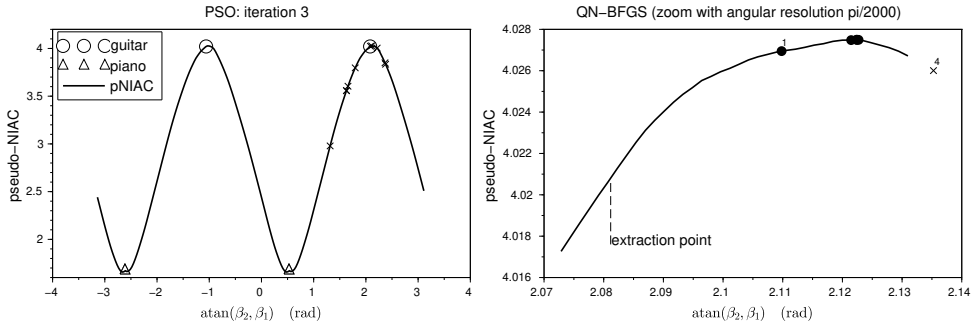



FIGURE 5.6 – Processus d'optimisation pour le mélange de 2 sources obtenu par extraction de la guitare du mélange de trois sources de la figure 5.5, puis déflation. À gauche, pseudo-NIAC en fonction de $\arctan(\beta_2, \beta_1)$, points d'extraction et essaim de particules à la dernière itération de la PSO. Les particules sont rassemblées autour du point maximisant la NIAC, correspondant à l'extraction de la guitare. À droite : zoom autour de ce maximum montrant les points successifs de l'optimisation QN-BFGS et le point d'extraction optimale.

| | SDR | SIR | SAR |
|---|---|---|---|
| voice | 49 | 49 | 73 |
| guitar | 28 | 28 | 73 |
| piano | 30 | 30 | 74 |

TABLE 5.1 – Pour un mélange déterminé de 3 sources, signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) et signal-to-artifact ratio (SAR) de la séparation de sources fondée sur la NIAC. Les sources sont ordonnées par ordre d'extraction.

### Étude des performances

Nous avons testé de manière plus systématique cette méthode de séparation de sources, avec le même paramétrage, sur un corpus de 5 extraits musicaux multi-pistes de la base QUASI [39, 27], de durées 9 à 16 s, ré-échantillonnés à 32 kHz.

Nous avons comparé les performances (en termes de SIR/SDR/SAR) avec celles de trois algorithmes de l'état de l'art fondés sur des critères différents :

— FastICA [19] maximise l'indépendance des sources extraites ;

— SOBI [3] utilise des statistiques du second ordre et une méthode de diagonalisation jointe [6] ;

— SEONS [9] repose sur les mêmes principes et tient compte de la non-stationnarité des sources.

Pour chaque nombre de sources $p = 3$ à 6 et pour chacun des 5 extraits, nous avons sélectionné 6 sources actives durant (presque) tout l'extrait et lancé les 4 méthodes de séparation pour chaque combinaison de $p$ sources parmi 6. Comme illustré par la figure 5.7, les méthodes ont des performances similaires, mais la proportion de SIR au-dessus de 40 dB est plus faible pour SOBI et SEONS, tandis que la proportion de SIR en-dessous de 20 dB est plus élevée pour la séparation fondée sur la NIAC. L'observation plus détaillée du processus d'optimisation pour ces cas conduit à trois types d'explication : (i) l'utilisation du critère d'indépendance évoqué en 5.3.2 pour choisir entre maximisation et minimisation conduit à un choix erroné ; (ii) la topographie de la fonction pseudo-NIAC à optimiser est difficile (par exemple un maximum sur une crête, entourée de quelques maxima locaux non-pertinents en

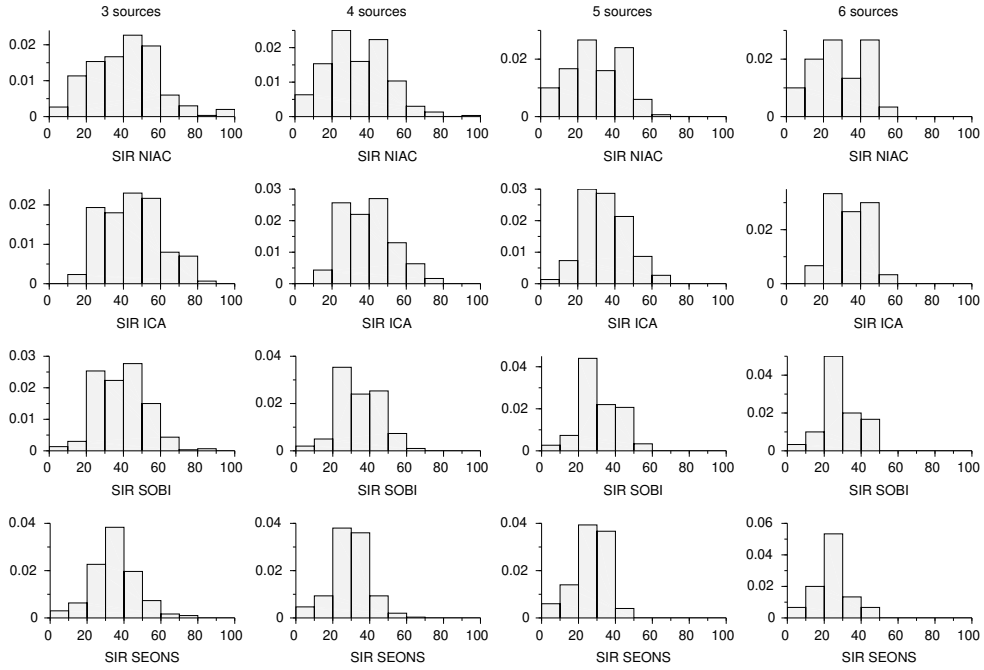



FIGURE 5.7 – Histogrammes des SIR résultant de la séparation de sources par FastICA, SOBI, SEONS et optimisation de la NIAC, pour des mélanges linéaires instantanés et déterminés de 3 à 6 sources.

termes de séparation) ; (iii) l'optimum au sens de la NIAC est légèrement différent de l'optimum au sens de la séparation (voir par exemple la figure 5.6).

Nous avons comparé la robustesse des différentes méthodes au mauvais conditionnement de la matrice de mélange. Notre méthode est légèrement plus robuste que FastICA, mais moins que SOBI et SOENS (voir détails dans [MSMS22]).

Dans ces expériences, FastICA a été utilisé en mode extraction-déflation, tandis que SOBI et SEONS estiment toutes les sources en même temps. Alors que l'ordre d'extraction dans FastICA dépend fortement de l'initialisation (aléatoire) de l'algorithme, ce qui produit une qualité d'estimation des sources très variable, les sources sont extraites dans un ordre assez stable en utilisant la NIAC, sauf pour des conditionnements sévères (de l'ordre de 1000 ou plus).

La séparation fondée sur la NIAC a l'avantage de ne pas reposer sur l'hypothèse d'indépendance des sources, contrairement aux méthodes du type ICA. L'indépendance est utilisé comme un critère secondaire pour choisir entre maximisation et minimisation, mais elle n'est pas fondamentale, de sorte que même des sources corrélées pourraient être séparées, un scénario où même SOBI et SEONS échouent.

Enfin, notre méthode de séparation ne nécessite pas l'hypothèse de non-gaussianité des sources, centrale dans les méthodes ICA. Nous avons testé notre méthode et FastICA sur le corpus précédemment évoqué en considérant toutes les mélanges de deux sources après gaussianisation de celles-ci par la méthode présentée au chapitre 3, sous-section 3.1.2. Alors que FastICA échoue ou atteint un SIR inférieur à 10 dB dans 77 % des cas, la séparation fondée sur la NIAC atteint un SIR moyen de 47 dB, avec 8 % des SIR inférieurs à 10 dB.

## 5.4 Conclusions

Nous avons montré qu'un bruit convolué par un signal audio permet de mesurer la netteté de celui-ci. Le bruit agit ici comme un révélateur du signal, une fonction qui peut être mise à profit pour extraire un son d'un mélange.

La NIAC présente plusieurs intérêts :

— elle est très corrélée au STI sans être intrusive ;

— contrairement aux autres méthodes non-intrusives, elle ne nécessite aucun apprentissage ou réglage fin de paramètres ;

— elle n'est pas spécifique à la parole ou la musique ;

— elle mesure une qualité intrinsèque du signal audio, indépendamment du canal de transmission ;

— la séparation de sources fondée sur la NIAC converge rapidement et se montre robuste à la gaussianité des sources, à leur dépendance, au mauvais conditionnement de la matrice de mélange et à l'initialisaton de l'algorithme.

Indépendamment des conditions d'exercice de la recherche, il aurait été logique de valider pleinement la NIAC comme mesure de netteté par une évaluation subjective formelle plus adaptée que celle présentée dans la sous-section 5.2.2, avant de l'étudier comme critère de netteté pour piloter des algorithmes de correction du son. Cette évaluation nécessite cependant des moyens logistiques et financiers qui étaient disponibles jusqu'en 2015, donc avant que la NIAC soit suffisamment développée, et nettement plus réduits après, notamment en 2020 et 2021 (restrictions sanitaires). D'autre part, il serait plus pertinent d'intégrer un modèle psychoacoustique dans la définition de la NIAC (qui repose sur une représentation temps-fréquence classique) avant d'étudier sa corrélation avec la netteté perçue. La définition d'un protocole de validation de la NIAC comme mesure de la netteté intrinsèque des sons (parole ou musique) reste néanmoins une des pistes de travail prioritaires.

Un autre facteur qui a conduit à écarter temporairement cette piste est le succès inattendu de l'utilisation de la NIAC comme critère de séparation de sources. La méthode présentée pour des mélanges linéaires instantanés et déterminées peut, en théorie, être facilement étendue à des mélanges convolutifs (ou à la déréverbération d'une source unique), mais l'expérimentation se heurte à des difficultés de stockage des variables intermédiaires (les $\Gamma_{X_i' X_j'}(f, f', \tau)$). Nous travaillons donc à son adaptation, ainsi qu'à l'adaptation aux mélanges instantanés sous-déterminés. Ces dernières années a émergé une autre application que nous souhaitons explorer, la restauration de la version originale (ou d'une version « propre ») d'un son à partir d'une multitude de versions « sales » (bruitées, réverbérées, saturées, compressées...) trouvées sur le web. Le caractère très rudimentaire des méthodes proposées en 2017 [34] laissait de la place à de nouvelles propositions.

Par sa définition, la NIAC n'est pas spécifique à l'audio : n'importe quel signal dont la pureté est caractérisée par la parcimonie de son spectrogramme pourrait bénéficier de cette approche, que ce soit pour l'évaluation de sa qualité ou pour des traitements de correction, notamment la séparation de sources.

Deux projets démarrent, dans lesquels la NIAC sera testée :

— Le projet ParkImVox a pour objet la détection précoce de la maladie de Parkinson et la discrimination Parkinson / Pakinson+, en assistant le diagnostic

par l'analyse de la voix, de l'électro-encéphalogramme (EEG) et d'IRM. La NIAC pourra être l'un des paramètres acoustiques discriminant.

— Le projet ÉPOPÉES XX-XXI vise notamment à archiver et éditorialiser les lectures et performances poétiques. Les enregistrements étant de qualité variable, l'étiquetage des entrées de la base devrait inclure la netteté perçue par les personnes chargées de cette tâche, ce qui peut constituer un moyen de validation de la NIAC (avec toutefois une grande variabilité des mesures subjectives due au peu de contrôle des conditions de test) permettant ultérieurement d'utiliser celle-ci comme outil d'étiquetage automatique.

## 5.5 Références bibliographiques


[1] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen. Nonintrusive speech intelligibility prediction using convolutional neural networks. *IEEE Trans. on Audio, Speech, and Language Processing*, 26(10) :1925–1939, 2018.

[2] ANSI. Methods for calculation of the speech intelligibility index. S3.5-1997.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2) :434–444, 1997.

[4] G. Blanchet and L. Moisan. An explicit sharpness index related to global phase coherence. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1065–1068, 2012.

[5] G. Blanchet, L. Moisan, and B. Rougé. Measuring the global phase coherence of an image. In *IEEE Int. Conf. on Image Processing*, pages 1176–1179, Oct 2008.

[6] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1) :161–164, Jan. 1996.

[7] P. Chavasse. De la notion d'intelligibilité. *L'audioprothésiste français*, 3, 1962.

[8] S. Cheng, Q. Qin, Z. Wu, Y. Shi, and Q. Zhang. Multimodal optimization using particle swarm optimization algorithms : CEC 2015 competition on single objective multi-niche optimization. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1075–1082, 2015.

[9] S. Choi, A. Cichocki, and A. Beloucharni. Second order nonstationary source separation. *J. VLSI Signal Process. Syst.*, 32(1–2) :93–104, Aug 2002.

[10] P. Comon and P. Jutten. *Handbook of Blind Source Separation.* Academic Press, 2010.

[11] G. D. Paced auditory serial-addition task : a measure of recovery from concussion. *Perceptual and motor skills*, 44(2) :367–373, 1977.

[12] V. D. Delmotte. *Computational Auditory Saliency.* PhD thesis, Georgia Institute of Technology, 2012.

[13] V. Duangudom and D. V. Anderson. Using auditory saliency to understand complex auditory scenes. In *2007 15th European Signal Processing Conference*, pages 1206–1210, 2007.

[14] T. H. Falk, C. Zheng, and W. Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(7) :1766–1774, Sept 2010.

[15] D. H. Griesinger. What is "clarity", and how it can be measured ? *Proc. of Meetings on Acoustics*, 19(1) :015003, 2013.

[16] P. H. Dual-task interference in simple tasks : data and theory. *Psychological bulletin*, 116(2) :220–244, 1994.

[17] M. E. Hossain, W. A. Jassim, and M. S. A. Zilany. Reference-free assessment of speech intelligibility using bispectrum of an auditory neurogram. *PLoS ONE*, 11(3), March 2016.

[18] T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. of America*, 77(3) :1069–1077, 1985.

[19] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3) :626–634, 1999.

[20] International Organization for Standardization. Acoustics - Measurement of room acoustic parameters - Part 1 : Performance spaces. ISO 3382-1 :2009.

[21] J. Jensen and C. H. Taal. Speech intelligibility prediction based on mutual information. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 22(2) :430–440, Feb 2014.

[22] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis. Mechanisms for allocating auditory attention : an auditory saliency map. *Current biology*, 15(21) :1943–1947, 2005.

[23] J. Kennedy. Particle swarm optimization. *Encyclopedia of machine learning*, pages 760–766, 2010.

[24] A. Lacheret, N. Obin, and X. Rodet. Un modèle de durées des syllabes fondé sur leurs propriétés intrinsèques et les variations locales de débit. In *Journées d'Etude sur la parole*, Avignon, France, 2008.

[25] A. Leclaire and L. Moisan. No-reference image quality assessment and blind deblurring with sharpness metrics exploiting Fourier phase information. *J. of Mathematical Imaging and Vision*, 52(1) :145–172, 2015.

[26] D. Lee, J. van Dorp Schuitman, X. Qiu, and I. Burnett. Development of a clarity parameter using a time-varying loudness model. *J. Acoust. Soc. of America*, 143(6) :3455–3459, 2018.

[27] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Trans. on Sig. Proc.*, 59 :3155–3167, 2011.

[28] N. Mechergui, S. Djaziri-Larbi, and M. Jaïdane. Speech based transmission index for all : An intelligibility metric for variable hearing ability. *The Journal of the Acoustical Society of America*, 141(3) :1470–1480, 2017.

[29] T. T. N. A comprehensive review of the paced auditory serial addition test (PASAT). *Archives of clinical neuropsychology*, 21(1) :53–76, 2006.

[30] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proc. of the IEEE*, 69(5) :529–541, May 1981.

[31] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. v. Waterschoot, and P. A. Naylor. A single-channel non-intrusive C50 estimator correlated with speech recognition performance. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(4) :719–732, April 2016.

[32] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes. A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*, 80 :84–94, 2016.

[33] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. of America*, 67(1) :318–326, 1980.

[34] N. Stefanakis, M. Viskadouros, and A. Mouchtaris. A subjective evaluation on mixtures of crowdsourced audio recordings. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1819–1823, 2017.

[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. on Audio, Speech, and Language Processing*, 19(7) :2125–2136, Sept 2011.

[36] F. Tordini, A. S. Bregman, J. R. Cooperstock, A. Ankolekar, and T. Sandholm. Toward an improved model of auditory saliency. In *19th International Conference on Auditory Display*, pages 189–196, Lodz, Poland, 2013.

[37] T. Tsuchida and G. W. Cottrell. Auditory saliency using natural statistics. In *Annual meeting of the cognitive science society*, pages 1048–1053, Sapporo, Japan, 2012.

[38] J. van Dorp Schuitman, D. de Vries, and A. Lindau. Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. *J. Acoust. Soc. of America*, 133(3) :1572–1585, 2013.

[39] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. Duong. The signal separation evaluation campaign (2007-2010) : Achievements and remaining challenges. *Signal Processing*, 92 :1928–1936, 2012.

[40] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469, 2006.

# Conclusion

Nous avons montré comment des distorsions inaudibles du son peuvent faciliter des traitements d'analyse ou de correction de celui-ci. À partir de la notion de tatouage audio, nous avons donné au bruit quatre fonctions au service du traitement du son : *informer* un système de traitement, *doper* le signal pour lui donner des propriétés adaptées au traitement visé ; *témoigner* des altérations subies par le signal pour les corriger ; *révéler* le signal. Cette démarche globale a été mise en œuvre dans plusieurs sous-domaines du traitement du signal audio : tatouage, compression, séparation de sources, identification de systèmes, quantification, codage de canal, qualité audio... Cette approche présente l'intérêt de déplacer la complexité algorithmique des traitements correctifs dans la chaîne de communication : le tatouage du signal en amont permet de rendre performants des algorithmes simples de traitement en aval.

Nous avons validé les idées fondatrices du projet WaRRIS, tout en les étendant. Cette validation a produit des solutions efficaces, notamment la correction des défauts des codecs, l'identification de systèmes, la mesure de netteté et la séparation de sources. Plusieurs sujets m'apportent la satisfaction du bel ouvrage :

— les thèses d'Imen Mezghani (identification de systèmes) et d'Imen Samaali (correction des codecs), par la richesse du travail accompli ;

— le théorème de quantification, par la joliesse des résultats et leur portée envisageable ;

— la NIAC, par la netteté de sa conception et son efficacité.

De nombreuses pistes ont été ouvertes, de manière arborescente, ce qui nécessite un élagage, qui ne concerne pas seulement les pistes sans résultats satisfaisants. Par exemple, le codage audio est un domaine très concurrentiel, où les travaux en cours ne sont pas toujours publiés, de sorte que toute solution peut se retrouver dépassée avant même sa publication. L'identification de système, malgré les bons résultats obtenus, est aussi un sujet que je ne souhaite pas poursuivre : en l'absence d'une application industrielle, il est dépourvu de beauté théorique et d'attrait ludique.

Les sujets suivants méritent d'être approfondis :

— le **tatouage par QIM** avec reformage spectral du bruit de quantification, eu égard à l'efficacité de la QIM et à ses bases théoriques solides

— le **reformage de densité de probabilité** au fil de l'eau. Toutes les applications de l'égalisation d'histogramme présentées supposent de traiter un enregistrement du signal ou de traiter le signal au fil de l'eau mais par blocs de longue durée (de l'ordre de la seconde), incompatibles avec les contraintes temps-réel de certaines applications. Il serait intéressant de considérer non pas l'histogramme mais une estimation de la densité de probabilité actualisée à

chaque nouvel échantillon (ou à chaque nouveau bloc d'analyse fréquentielle, dans le cas d'une analyse temps-fréquence).

— le **théorème de sous-quantification** étendu à la densité de probabilité jointe de plusieurs échantillons successifs. En considérant l'histogramme correspondant pour une séquence finie, reconstruire l'histogramme d'un signal quantifié à partir de sa version sous-quantifiée permettrait de retrouver le signal quantifié grâce à la limitation des combinaisons possibles.

— le **tatouage témoin** dans le cas, à définir, de signaux représentables dans un espace où le tatouage distordu est séparable du reste du signal.

— la validation de la **NIAC** comme mesure de netteté intrinsèque par des tests subjectifs formels, son utilisation pour séparer des mélanges convolutifs ou sous-déterminés, son utilisation pour reconstruire un signal à partir de multiples versions dégradées et son extension au-delà de l'audio.

Cet ensemble consiste à la fois à renforcer les fondements (outils théoriques et pratiques) de l'approche présentée dans ce mémoire et à approfondir certaines applications qui nous semblent avoir une grande portée.

Au-delà d'un programme technique précis, la suite du travail est aussi de continuer à user de la liberté du chercheur [4] pour se concentrer sur les aspects ludiques du traitement du signal et sur la beauté de ses constructions théoriques, indépendamment des modes et du conditionnement des moyens de travail à la réponse à des *« défis sociétaux »* [5]. Il ne s'agit ni d'une profession de foi en futilité ni de refuser le principe d'une utilité sociale du chercheur, bien au contraire.

D'une part, le pilotage de la recherche par des *« demandes de la société »* [6] néglige la capacité d'une recherche non-orientée à produire des résultats utiles et soulève de nombreuses questions sur les approches technocentrées qui dominent en traitement du signal, pour des sujets qui relèvent d'abord de l'organisation politique et sociale. Si l'on s'accorde sur l'inutilité sociale d'une recherche visant à faire communiquer la cafetière et le frigo, le chercheur en traitement du signal participant à la conception de solutions de télé-médecine pour compenser la désertification médicale... œuvre-t-il avec une utilité sociale ou au service du *social washing* ?

D'autre part, l'utilité sociale doit se manifester avec force dans le lien enseignement-recherche. Intéresser les étudiants au traitement du signal, discipline souvent considérée comme rébarbative, peut passer par la réponse à la question *« À quoi ça sert ? »* ; les applications concrètes sont alors une récompense stimulante pour les heures passées devant des équations. Mais avant tout, la capacité de l'enseignant-chercheur à partager certains aspects ludiques ou à faire apprécier la beauté de constructions théoriques relève de l'essence même de l'utilité sociale de notre métier.

4. renforcée par la faible dépendance de cette recherche aux moyens matériels
5. pour reprendre le vocable de l'ANR
6. dont le mode de recensement reste flou

# Liste des publications

Les noms des étudiant.e.s encadré.e.s sont soulignés. Toutes les publications sont accessibles sur `https://helios2.mi.parisdescartes.fr/~mahe/Recherche/HDR/publis.html`, sauf l'ouvrage pédagogique.

**Revues internationales avec comité de lecture et sélection sur article long**

[1] **Gaël Mahé**, André Gilloire, and Lætitia Gros. Correction of the voice timbre distortions in telephone networks : method and evaluation. *Speech Communication*, 43(3) :241–266, August 2004.

[2] Imen Samaali, **Gaël Mahé**, and Monia Turki. Watermark-aided pre-echo reduction in low bit-rate audio coding. *J. Audio Eng. Soc*, 60(6) :431–443, june 2012.

[3] **Gaël Mahé**, Everton Nadalìn, Ricardo Suyama, and João Romano. Perceptually controlled doping for audio source separation. *EURASIP Journal on Advances in Signal Processing*, 2014(1) :27, march 2014.

[4] I. Mezghani-Marrakchi, **G. Mahé**, S. Djaziri-Larbi, M. Jaidane, and M. Turki-Hadj Alouane. Nonlinear audio systems identification through audio input gaussianization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1) :41–53, january 2014.

[5] S. Djaziri-Larbi, **G. Mahé**, I. Mezghani, M. Turki, and M. Jaïdane. Watermark-driven acoustic echo cancellation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2) :367–378, Feb 2018.

[6] **Gaël Mahé** and Mériem Jaidane. Perceptually Controlled Reshaping of Sound Histograms. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(9) :1671 – 1683, September 2018.

[7] Danping Wang, **Gaël Mahé**, Junying Fang, Julien Piscione, Serge Couvet, Didier Retière, Sébastien Laporte, and Pierre-Paul Vidal. Inconsistent anticipatory postural adjustments (APAs) in rugby players : a source of injuries ? *BMJ Open Sport & Exercise Medicine*, 4(1), 2018.

[8] Danping Wang, **Mahé, Gaël**, Junying Fang, Julien Piscione, Serge Couvet, Didier Retière, Sébastien Laporte, and Pierre-Paul Vidal. Collaborative sensorimotor intelligence : the scrum as a model. *BMJ Open Sport & Exercise Medicine*, 4(1), 2018.

[9] **Gaël Mahé**, Giulio G.R. Suzumura, Lionel Moisan, and Ricardo Suyama. A non intrusive audio clarity index (NIAC) and its application to blind source separation. *Signal Processing*, 194 :108448, 2022.

**Papier invité court dans une conférence internationale**

[10] Djaziri-Larbi S., **Mahé G.**, Mezghani-Marrakchi I., Turki-Hadj-Allouane M., and Jaïdane-Saïdane M. Doping and witness watermarking for audio processing. In *Proc. of the 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA 2011)*, Tipaza, Algeria, may 2011.

**Communications internationales avec actes, comité de lecture et sélection sur article court**

[11] **G. Mahé** and A. Gilloire. Correction of the voice timbre distortions on telephone network. In *Proc. Eurospeech 2001*, pages 1867–1870, Aalborg, Denmark, september 2001.

[12] **G. Mahé** and A. Gilloire. Quantization noise spectral shaping in instantaneous coding of spectrally unbalanced speech signals. In *Proc. IEEE Workshop on Speech Coding*, pages 56–58, Tsukuba, Ibaraki, Japon, October 2002.

[13] **G. Mahé** and A. Gilloire. Multi-referenced correction of the voice timbre distortions on telephone network. In *Proc. Eurospeech 2003*, pages 1381–1384, Geneva, Switzerland, september 2003.

[14] Aline Neves, **Gaël Mahé**, and Mamadou Mboup. Restoration of voice timbre in telephone networks, based on both voice and lines properties. In *Proceedings of the 12th European Signal Processing Conference (Eusipco 2004)*, pages 1943–1946, Vienna, Austria, 2004.

[15] I. Marrakchi, M. Turki-Hadj Alouane, S. Djaziri-Larbi, M. Jaïdane-Saïdane, and **G. Mahé**. Speech processing in the watermarked domain : Application in adaptive echo cancellation. In *Proceedings of the 14th European Signal Processing Conference (Eusipco 2006)*, Florence, Italy, 2006.

[16] I. Marrakchi, **G. Mahé**, M. Jaidane-Saidane, S. Djaziri-Larbi, and M. Turki-Hadj Alouane. Gaussianisation method for identification of memoryless nonlinear audio systems. In *Proceedings of the 15th European Signal Processing Conference (Eusipco 2007)*, pages 2316–2320, 2007.

[17] I. Samaali, **G. Mahé**, and M. Turki-Hadj Alouane. Criteria to measure the quality of TVAR estimation for audio signals. In *Proceedings of the 15th European Signal Processing Conference (Eusipco 2007)*, pages 798–802, Poznan, Poland, 2007.

[18] H. Halalchi, **G. Mahé**, and M. Jaïdane. Revisiting quantization theorem through audiowatermarking. In *Proc. ICASSP 2009*, pages 3361–3364, Taipei, Taïwan, 2009.

[19] I. Samaali, M. Turki-Hadj Alouane, and **G. Mahé**. Temporal envelope correction for attack restoration in low bit-rate audio coding. In *Proceedings of the 17th European Signal Processing Conference (Eusipco 2009)*, pages 929–933, Glasgow, United Kingdom, august 2009.

[20] I. Samaali, M. Turki-Hadj Alouane, and **G. Mahé**. Attack localization based on algebra detector for pre-echo reduction in low bit-rate audio coding. In *Proc. 5th International Symposium on Image/Video Communications and Mobile Networks (ISIVC)*, Rabat, Morocco, september 2010.

[21] **G. Mahé**, E.Z. Nadalin, and J.M.T. Romano. Doping audio signals for source separation. In *Proceedings of the 20th European Signal Processing Conference (Eusipco 2012)*, pages 2402–2406, Bucarest, Romania, August 2012.

[22] I. Samaali, **G. Mahé**, and M. T. H. Alouane. High-frequency tonal components restoration in low-bitrate audio coding using multiple spectral translations. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1053–1057, Aug 2015.

[23] **G. Mahé**, L. Moisan, and M. Mitrea. An image-inspired audio sharpness index. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 683–687, Aug 2017.

[24] F. El-Jili, **G. Mahé**, and M. Mboup. A robust signal quantization system based on error correcting codes. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2561–2565, Aug 2017.

**Communications francophones avec actes, comité de lecture et sélection sur article court**

[25] **Gaël Mahé** and André Gilloire. Contrôle de l'audibilité du bruit de quantification induit par la pré-distorsion d'un signal de parole. In *Proc. GRETSI*, pages 237–240, Paris, France, september 2003.

[26] Mezghani-Marrakchi I., Turki-Hadj-Allouane M., Djaziri-Larbi S., Jaïdane-Saïdane M., and **Mahé G.** Analyse des performances d'une nouvelle structure d'aec dans le domaine tatoué. In *Proc. of International Symposium on Image/Video Communications (ISIVC'06)*, Hammamet, Tunisia, september 2006.

[27] Tifanie Bouchara and **Gaël Mahé**. Evaluation de la saillance d'annonces vocales par un paradigme de double-tâche. In *Actes du 12ème Congrès Français d'Acoustique (CFA2014)*, pages 625–631, Poitiers, France, april 2014.

**Brevets**

[28] **Gaël Mahé** and André Gilloire. Procédé et dispositif de correction centralisée du timbre de la parole sur un réseau de communications téléphoniques. Brevet déposé en France le 28.03.01 (no de dépôt FR0104194), 2001.

[29] **Gaël Mahé** and André Gilloire. Procédé et système de correction multi-références des déformations spectrales de la voix introduites par un réseau de communication. Brevet déposé le 11.12.02 (no de dépôt FR0215618, étendu à l'Europe sous le no EP1429316 et aux USA sous le no US20040172241), 2003.

[30] Mezghani-Marrakchi I., **Mahé G.**, Jaïdane-Saïdane M., Djaziri-Larbi S., and Turki-Hadj-Allouane M. Procédé et dispositif d'annulation d'écho acoustique par tatouage audio. brevet déposé en France le 29 octobre 2009 sous le no 0957636, étendu à l'international sous le no WO/2011/051625, 2009.

**Ouvrage pédagogique**

[31] **Gaël Mahé**. *Systèmes de communications numériques*. Ellipses, Paris, 2012. cours et exercices corrigés.

**En cours de soumission**

[32] **G. Mahé**, F. El-Jili, and M. Mboup. A robust signal quantization based on error correcting codes. soumis au *Journal on Advances on Signal Processing*.

# Liste des acronymes d'institutions

| | |
|---|---|
| AAL | Ambient Assisted Living |
| ANR | Agence Nationale de la Recherche |
| AUF | Agence Universitaire de la Francophonie |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| CEA | Commissariat à l'Énergie Atomique et aux énergies alternatives |
| CECS | Centro de Engenhara, modelagem e Ciências Sociais aplicadas |
| CMCU | Comité Mixte franco-tunisien de Coopération Universitaire |
| COFECUB | Comité Français d'Évaluation de la Coopération Universitaire et scientifique avec le Brésil |
| CRESTIC | Centre de Recherche en Sciences et Technologies de l'Information et de la Communication (Université de Reims) |
| DSPcom | Digital Signal Processing and Communications (Unicamp) |
| ENEA | Agenzia Nazionale per le Nuove tecnologie, l'Energia e lo Sviluppo economico sostenibile (Italie) |
| ENIT | École Nationale d'Ingénieurs de Tunis |
| ENST | École Nationale Supérieure des Télécommunications |
| IRCAM | Institut de Recherche en Coordination Acoustique et Musique |
| IRENav | Institut de recherche de l'École navale |
| L3S | Laboratoire Signals and Smart Systems (ex-U2S, ENIT) |
| LIPADE | Laboratoire d'Informatique Paris Descartes (ex-CRIP5, UPC) |
| LTCI | Laboratoire Traitement et Communication de l'Information |
| MAP5 | laboratoire Mathématiques Appliquées à Paris 5 (UPC) |
| SATIE | Systèmes et Applications des Technologies de l'Information et de l'Énergie (ENS Paris Saclay) |
| SupCom | École supérieure des communications de Tunis |
| UFABC | Universidade Federal do ABC (Brésil) |

| | |
|---|---|
| U2S | Unité Signaux et Systèmes (ENIT) |
| Unicamp | Universidade Estadual de Campinas (Brésil) |
| UPC | Université Paris Cité (ex-Université de Paris, ex-universités Paris Descartes (Paris 5) et Paris Diderot (Paris 7)) |

# Projets et financements

*Plus de détails sur les projets sont donnés dans le dossier d'accompagnement. Les acronymes sont développés page 121.*

**[CMCU2004]** Projet CMCU 04G0201 *Acoustique et Musique : Nouvelles Techniques Numériques de Restitution Sonore* (2004-2007) avec l'U2S, l'IRCAM et l'Institut Supérieur de Musique de Sousse. L'objectif est d'initier en Tunisie, à travers une collaboration avec des partenaires français, une structure de réflexion multidisciplinaire, dans le domaine de la recherche-développement, en acoustique et musique. L'accent est mis sur la notion de mesure temps réel de la qualité acoustique d'une salle dans un esprit d'analyse (lors d'un enregistrement dans un lieu donné), de conception (pour l'amélioration temps réel de la qualité audio par des techniques d'annulation d'écho acoustique, de restauration de timbre,...) et de synthèse (pour la déréverbération, la compensation de contexte en spatialisation du son,...).

**[WaRRIS]** Le projet ANR jeunes chercheurs ANR-06-JCJC-0009 (*Watermarking Réflexif pour le Renforcement des Images et des Sons (WaRRIS, 2006-2010)* avec le MAP5 et l'U2S, vise à définir des tatouages audio et image intimement liés au signal hôte, contenant des informations utiles à la restauration, l'analyse ou la manipulation de ce signal à la sortie d'un canal de communication dégradé.

**[AUF]** Bourse de mobilité de l'Agence Universitaire de la Francophonie (AUF) : séjour de 3 mois à l'U2S en 2007.

**[EReQCA]** Projet CMCU 08S1414 *Évaluation et renforcement de la qualité en communication audio* (2008-2011) avec l'U2S (ENIT), l'IRENav (École Navale), Techtra (SupCom) et l'opérateur télécom Tunisiana. Ce projet a pour thème l'évaluation et l'amélioration de la qualité du signal audio obtenu en sortie des nouveaux systèmes de communications numériques (Voix sur IP, transmissions radio-mobiles 3ème génération, radio mondiale, ...). Il vise des méthodes objectives et subjectives d'évaluation et des méthodes de codage et de débruitage fondées sur des approches perceptives.

**[ICityForAll]** Projet européen AAL 2011-4-056 ICityForAll (la ville intelligible pour tous / les TIC pour tous) du programme européen AAL (2012-2015) avec CEA, Centre d'Expertise National des Technologies de l'Information et de la Communication pour l'autonomie (CENTICH), Active Audio, École Polytechnique Fédérale de Lausanne, Université Technique de Munich, FIAT, ENEA,Telnet (Tunisie),

Université de Reims, European Social Cooperative (ESCOOP, Italie). L'objectif est d'améliorer l'intelligibilité des signes sonores dans les espaces urbains. Les routes, gares, aéroports, centres commerciaux, halls d'accueil... sont en effet largement organisés *via* des annonces et indications sonores. La compréhension des annonces est perturbée par le bruit ambiant et la réverbération, notamment pour les personnes malentendantes. La perception, l'identification et la localisation des alarmes sont également difficiles pour les malentendants.

**[Compest]** Projet CAPES-COFECUB *Estimation paramétrique robuste de systèmes linéaires sous-déterminés par la technique du compressive sensing* (Compest) (2013-2017) avec SATIE et DSPCom. Les applications visées sont l'identification parcimonieuse de canaux de propagation radiomobile, la séparation de sources et l'estimation parcimonieuse d'un système linéaire pour son asservissement en automatique.

**[ROAPI]** Projet CAPES-COFECUB *Représentations, optimisations et algorithmes parcimonieux pour problèmes inverses mal posés* (ROAPI) (2020-2023) avec SATIE et DSPCom. L'objectif est d'améliorer la parcimonie des modèles utilisés dans les problèmes inverses pour les adapter à des algorithmes fondés sur l'hypothèse de parcimonie et, réciproquement, d'adapter ces algorithmes aux représentations obtenues. Les champs d'applications sont notamment la géolocalisation et la séparation de sources audio.

**[ParkImVox]** Projet CMCU 22G1117 ParkImVox (2022-2024) avec le L3S, le Laboratoire des Maladies Neurodégénératives et troubles Psycho-comportementaux (LR18SP03) du CHU Razi de Tunis et le Laboratoire Maladies Neurologiques de l'Enfant : Investigations et Prise en Charge (LR18SP04) de la Faculté de Médecine de Tunis. Ce projet vise la détection précoce de la maladie de Parkinson et la discrimination Parkinson / Pakinson+, en assistant le diagnostic par l'analyse de la voix, de l'électro-encéphalogramme et d'IRM.

**[ÉPOPÉES XX-XXI]** Le projet ÉPOPÉES XX-XXI [Écrire l'histoire des POésies PerformÉES aux XXe et XXIe siècles], sélectionné pour la deuxième phase de soumission à l'appel à projet ANR générique 2022, vise à archiver et éditorialiser les lectures et performances poétiques, ainsi qu'à repenser le genre poétique par ses manifestations orales. Il est porté par le Laboratoire de Recherche sur les Cultures Anglophones (LARCA-UMR8225, Université Paris Cité), en partenariat avec Sorbonne Université, l'Université Jean Moulin-Lyon 3, CY Cergy Paris Université et l'École Normale Supérieure.

# Remerciements

Si la fusion entre les universités Descartes et Diderot a eu un effet positif, c'est de me permettre de proposer à Laurent Daudet d'être le représentant de mon université dans le jury de cette habilitation. Laurent a accepté de présider celui-ci ; je l'en remercie. Merci à Sylvain Marchand, Phillip A. Regalia et Gaël Richard pour leur travail de rapporteur. Leurs analyses complémentaires ont éclairé mon travail d'une manière souvent originale, qui sera profitable pour la suite. Je remercie Rosângela Coelho, Régine Le Bouquin Jeannès, Mériem Jaïdane et Olivier Warusfel d'avoir participé au jury. Plusieurs de leurs questions ont ouvert des pistes de réflexion inspirantes. Plus généralement, les membres de jury ont fait de cette soutenance un moment d'échange scientifique riche et cordial.

Cette histoire a commencé avec André Gilloire, qui a encadré ma thèse au CNET, devenu Orange Labs. Je lui suis reconnaissant de la confiance, de l'intérêt et de la gentillesse qu'il y a mis, ainsi que de m'avoir transmis (ou favorisé ce qui était en germe ?) ce que je crois être sa conception de la recherche : une recherche libre, alliant rigueur et créativité et se nourrissant de notre capacité à nous enthousiasmer.

J'ai retrouvé cet esprit en arrivant à l'Université Paris Descartes, auprès de Madeleine Bonnet, puis de Mériem Jaïdane. Un grand merci à Madeleine pour m'avoir tant soutenu dans cette entrée dans le monde universitaire que je ne connaissais pas, aux mœurs parfois étranges. Alors que nous avons fait peu de recherches ensemble, une grande part de mes travaux te doit beaucoup, car réalisés avec les personnes clés que j'ai rencontrées grâce à toi : Mamadou, Mériem et João-Marcos.

Avant de partir à Reims, Mamadou Mboup était l'indispensable voisin avec qui partager un problème de traitement du signal. Depuis, je ne fais plus autant appel à ses lumières, mais j'apprécie toujours son flegme et son humour acéré face à la médiocrité.

Si je me suis efforcé de retracer dans ce mémoire la genèse des idées qui y sont présentées, il s'agit d'une reconstruction *a posteriori*, nécessairement lissée et incomplète. Pour une partie de ces travaux, reconstituer précisément, à la manière d'un archéologue, le cheminement d'une idée à l'autre ou d'un résultat à une idée, aurait nécessité d'enregistrer les longues et foisonnantes discussions avec Mériem Jaïdane, où se mêlent le scientifique et le reste, le farfelu et le théorique très sérieux. Merci Mériem pour ces bons moments et pour ton talent d'accoucheuse d'idées.

Grâce à Mériem, j'ai eu la chance de connaître deux autres piliers de l'Unité Signaux et Systèmes (U2S) de Tunis, Monia et Sonia. Quand Monia Turki m'a proposé de travailler sur les modèles TVAR, je n'ai pas osé lui dire que ça n'avait pas l'air très amusant, et j'ai bien fait, car ce premier travail ensemble a été le début d'une longue et fructueuse collaboration. Sonia Larbi m'a non seulement fait découvrir de

très bons petits restaurants, mais elle est surtout une partenaire de recherche rigoureuse et efficace, avec qui j'ai été très heureux de réaliser une partie de ces travaux. Avec Mériem, Monia et Sonia, j'ai co-encadré les thèses de Imen Mezghani et Imen Samaali : Imen et Imen ont fait un travail remarquable et mon cheminement en recherche leur doit beaucoup.

À l'U2S, j'ai aussi fait la connaissance de Sylvie Sevestre. Son énergie communicative a lancé et fait vivre nombre de projets franco-tunisiens, dont le projet ICity-ForAll, qui a suscité une partie des travaux présentés ici. Dans ce projet, j'ai eu le plaisir de travailler avec Tifanie Bouchara, dont les qualités scientifiques et humaines nous ont aidés à dépasser notre inexpérience en matière de mesure de netteté audio.

En 2011, João-Marcos Romano m'a aidé à mettre un pied dans la séparation de sources et hors de mon labo. Grâce à lui, j'ai rencontré plusieurs des chercheurs brésiliens qu'il sait fédérer dans une ambiance chaleureuse : c'est ainsi que j'ai eu la chance de travailler avec Everton Nadalin, Ricardo Suyama, Giulio Suzumura, Leonardo Tomazeli Duarte et Kazuo Takahata.

Sans aller si loin, à quelques portes de mon bureau, d'autres collègues m'ont apporté un peu d'air dans ce début des années 2010 tendu. Danping Wang a eu la bonne idée de me proposer de mettre le traitement du signal au service du rugby et a confirmé, par sa personnalité, que les gens avec qui nous travaillons sont au moins aussi importants que le sujet. Lionel Moisan m'a fait la proposition géniale de transposer à l'audio sa mesure de netteté d'image, le sharpness index. Je me suis alors engagé dans de longs tâtonnements dont a émergé la NIAC, aux résultats si prometteurs. Qu'il en soit remercié.

Hors de ces collaborations directes en recherche, il y a de nombreux collègues — enseignant·es-chercheur·es, chercheur·es, administratif·ves, ingénieurs et techniciens — dont j'estime le travail et avec qui j'ai aimé échanger. Ils et elles ont fait de mon UFR un environnement où l'on se sent bien.

Ce travail doit aussi beaucoup à l'énergie que donne la lutte quand on est épaulé par un collectif fort : merci aux camarades de la FSU et du collectif PUPH pour leur engagement pour une université où chacun·e puisse se réaliser dans son travail.

Merci enfin aux ami·es qui ont accompagné ces années. Une pensée particulière pour Robert et Muriel, qui ont facilité le début et la fin de cette histoire, ou plutôt de ce long épisode : c'est grâce à eux que j'ai sauvé l'achèvement de mon doctorat d'un problème sur mon contrat d'ATER ; c'est chez eux, à Tréveneuc, face à la mer, que j'ai relu et corrigé tranquillement ce mémoire.