INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° : ␣␣␣␣␣␣␣␣␣␣

THÈSE

pour obtenir le grade de

DOCTEUR de l'INP Grenoble

Spécialité : **Signal, Image, Parole, Télécoms**

préparée aux laboratoires
Institut de la Communication Parlée, UMR CNRS 5009
Laboratoire des Images et des Signaux, UMR CNRS 5083

dans le cadre de l'Ecole Doctorale
« **Électronique, Électrotechnique, Automatique et Traitement du Signal** »

présentée et soutenue publiquement par

Bertrand Rivet

le 29 septembre 2006

La bimodalité de la parole au secours de la séparation de sources

Directeur de thèse : Christian Jutten (LIS)
Co-directeur de thèse : Laurent Girin (ICP)

JURY

M.	Éric Moulines,	Président
M.	Yannick Deville,	Rapporteur
Ms.	Frédéric Bimbot / Rémi Gribonval,	Rapporteurs
M.	Christian Jutten,	Directeur de thèse
M.	Laurent Girin,	Co-directeur de thèse
M.	Dinh-Tuan Pham,	Éxaminateur

"Rien ne va de soi. Rien n'est donné. Tout est construit"
Gaston BACHELARD, *La Formation de l'esprit scientifique*, 1938.

Remerciements

Avant toute chose, je tiens à remercier Jean-Luc Schwartz et Jean-Marc Chassery directeurs respectifs de l'Institut de la Communication Parlée (ICP) et du Laboratoire des Images et des Signaux (LIS) pour m'avoir accueilli dans leurs laboratoires.

Je veux particulièrement remercier mes deux papas de thèse Christian et Laurent : sans vous cette thèse n'aurait jamais vu le jour. Vous avez su m'encadrer sans être directifs et me communiquer l'énergie nécessaire pour mener à bien mes recherches. Même lorsque le temps m'aura pressé, vous avez toujours su trouver un moment pour relire de mon manuscrit. Si l'envie de continuer de vous remercier ne manque pas, les mots justes sont plus difficiles à trouver et finalement les plus simples sont sûrement les plus vrais : encore un grand merci. Je penserai à vous avec mon Chivas de 18 ans d'âge et mon stylo !

I would like to thank Jonathon Chambers, the leader of the Centre of Digital Signal Processing at Cardiff University for your care during my stay. I also would like to thank Yulia and Andrew. I was really glad by the collaboration. Thank you very much for your care.

Je souhaite associer à ces remerciements Patrice Petitclair pour avoir été mon tuteur pédagogique, pour m'avoir fait confiance en TP et me laisser encadrer des TDs, chose suffisamment rare pour que je prenne le temps de te remercier. Ces quelques heures d'enseignement auront été une bouffée d'air frais.

Je voudrais également remercier tous les membres de mon jury pour avoir pris le temps de lire et de critiquer ce manuscrit : Eric Moulines pour avoir accepté de présider ce jury, Yannick Deville, Frédéric Bimbot et Rémi Gribonval pour avoir apporté votre caution et vos remarques en rapportant mon travail et finalement Dinh-Tuan Pham pour ses précieuses suggestions.

Un grand merci à tous ceux que j'ai pu côtoyer et apprécier pendant ma thèse. A Claire, Julie, Antoine et Mohammad, mes compagnons de cordée à l'ICP. J'espère vous retrouver bientôt. Merci Anthony, mon colloc qui n'a jamais gueulé alors que tu as eu plus d'une occasion de le faire. Merci Jérémy pour avoir parfois délaissé ta douce pour faire des parties de billard. Claire, le pardonneras-tu ? Merci Annemie, tes mails, ta bonne humeur et ta présence sans faille ont été d'une fraîcheur très appréciable. Merci Claire pour toutes ces discussions pendant les formations obligatoires et pour tout ce que tu as fait pour moi. Merci Julie, ta gentillesse, ton rire et tes gâteaux ont toujours égayé plus que la pause café. Merci Antoine pour ces délires de fin de rédaction. Merci Guillaume, on aura bien rigolé pendant les pauses café. Merci Virginie, ta fraîcheur et ton attention m'auront bien aidé. Merci pour tout Nino, ou plutôt "Monsieur Medves" (j'ai failli râté ma sortie !). Merci Popo pour tes

remarques aussi spontanées qu'imprévisibles. Merci David pour tout à la fois nos remarques sur les enseignements et ces soirées en conférence. Enfin un grand merci à tous les membres de l'ICP et du LIS.

Je tiens aussi à remercier Alex, Mourroun, Sonic, Cédric et Gizmo pour ces escapades parisiennes.

Merci à toute la bande. Djey et Clara pour m'avoir présenté. Sab, Amande, Cécile, Serge, Djouls, Florent, Tons, Rob pour ces sorties skis ou ces soirées raclettes et tartiflettes. Merci Clochette pour tes histoires hors du commun. Merci "serruria", toi la danseuse slave, pour nos discussions. Mais il me faut maintenant te dire la vérité sur ton arbre généalogique scientifique : non Fant n'est pas ton grand-père et Chiba n'est pas ton arrière grand-père. Voila c'est fait, c'est sûrement dur et violent mais c'était nécessaire. De toute façon "j'chte déteste". Merci "ma maman préférée", ton "petit snorky" a maintenant fini. Merci GrG mon grand frère pour tout ce que l'on nous a attribué à tord ou à raison. Merci aux membres du comité de sélection de la Cherch'Ac qui n'aura pas survécu plus d'une saison. Enfin, un grand merci aux colocataires du bureau 523 pour les bonbons, vous m'avez sauvé d'atroces hypoglycémies.

Pour avoir un exemplaire unique de cette thèse, créez vos propres remerciements en complétant la liste ci-dessous.
Je voudrais remercier _____ pour sa contribution hors norme sans toi cette thèse n'aurait pas été intéressante.
Merci à toi _____, mon père spirituel / ma mère spirituelle[1].
Je te remercie _____ pour ta bonne humeur, tes blagues plus drôles les unes que les autres, tu m'auras fait rire même dans les moments difficiles.
Merci _____ pour ton immense culture et tes discussions d'un niveau intellectuel qui m'étaient jusqu'à présent inconnues.

[1] Rayer la mention inutile.

Table des matières

Notations mathématiques

Fonctions et opérateurs

$\mathcal{H}(\cdot)$	Processus de mélange
$\mathcal{G}(\cdot)$	Processus de séparation
$\det(\cdot)$	Déterminant d'une matrice
$\ln(\cdot)$	Logarithme népérien (ou naturel)
$\log(\cdot)$	Logarithme décimal
$\mathrm{TF}(\cdot)$	Opérateur transformée de Fourier
$(\cdot)'$	Dérivation
$\widehat{(.)}$	Estimée
$\lvert\cdot\rvert$	Module pour les scalaires
$\lvert\cdot\rvert$	Module composante par composante pour les vecteurs
$\lvert\cdot\rvert$	Cardinal pour un ensemble
$(\cdot)^*$	Conjugaison
$(\cdot)^T$	Transposition
$(\cdot)^+$	Transposition conjugaison
$(\cdot)^\dagger$	Pseudo-inverse d'une matrice rectangulaire
$(\cdot) * (\cdot)$	Produit de convolution
$(\cdot) \circ (\cdot)$	Composition
$(\cdot) \oplus (\cdot)$	Somme directe de deux espaces
$(\cdot) \overset{\perp}{\oplus} (\cdot)$	Somme directe orthogonale de deux espaces

Variable aléatoire

$\Pr[\cdot]$	Probabilité d'un événement
$p[\cdot]$	Densité de probabilité d'une variable aléatoire
$\mathrm{P}[\cdot]$	Fonction de répartition d'une variable aléatoire
$\Psi[\cdot]$	Fonction score d'une variable aléatoire
$\mathrm{E}[\cdot]$	Espérance mathématique
$\mathrm{Var}[\cdot]$	Variance mathématique
$H[\cdot]$	Entropie de Shannon
$I[\cdot]$	Information mutuelle
$KL[\cdot\|\cdot]$	Divergence de Kullback-Leibler
$\Phi[\cdot]$	Fonction de contraste
$\Phi^{\circ}[\cdot]$	Fonction de contraste orthogonal
$\mathcal{N}(\mu, \Gamma)$	Loi normale de vecteur moyenne μ et de matrice de covariance Γ

Ensembles et espaces

$\mathbb{R}$	Ensemble des réels
$\mathbb{C}$	Ensemble des complexes
$\{\cdot\}_i$	Ensemble formé des éléments dépendants de i pour tous les i
$\mathrm{U}(n)$	Groupe des matrices unitaires de taille $(n \times n)$
$\mathrm{D}(n)$	Ensemble des matrices diagonales de taille $n \times n$

Grandeurs scalaires

N_f	Nombre de fréquences de calcul des transformées de Fourier
N_s	Nombre de sources
N_o	Nombre d'observations

Signaux et grandeurs vectorielles

$\mathbf{y}(t)$	Vecteur colonne de signaux temporel
$y_i(t)$	$i^{\text{ème}}$ composante du vecteur $\mathbf{y}(t)$

Grandeurs matricielles

A	Matrice
$A_{i,j}$	$(i,j)^{\text{ème}}$ élément de la matrice A
I_n	Matrice identité de taille $(n \times n)$
Π	Matrice de permutation dont la taille est à préciser
$\Lambda(\cdot)$	Matrice diagonale de distorsion

Abréviations

Nous donnons entre parenthèses l'abréviation anglaise si elle est différente.

ACI	Analyse en composantes indépendantes (ICA)
ACP	Analyse en composantes principales (PCA)
COR	caractéristiques opérationnelles de réception
dB	Décibel
EASI	Algorithme équivariant adaptatif (Equivariant Adaptative Separation via Independence)
ECG	Electrocardiogramme
iid	indépendant et identiquement distribué
LP	Prédiction linéaire (Linear Prediction)
MMG	Modèle multi-gaussien (GMM)
MMLR	Modèle multi-LogRayleigh (LRMM)
RAP	Reconnaissance automatique de la parole (ASR)
RI	Réponse impulsionelle (IR)
RSB	Rapport signal sur bruit (SNR)
RSI	Rapport signal sur interférence (SIR)
SAS	Séparation aveugle de source (BSS)
SOBI	Identification aveugle au second ordre (Second Order Blind Identification)
TCD	Transformée en cosinus discrète (DCT)
TF	Transformée de Fourier (FT)
TFCT	Transformée de Fourier à court terme (STFT)
TFD	Transformée de Fourier discrète (DFT)

Introduction

Ne vous êtes-vous jamais demandé pourquoi nous avons tant de mal à entendre ce que dit le conducteur d'une voiture lorsque nous sommes à l'arrière, surtout si notre environnement devient très bruyant ? Nous savons que notre cerveau est capable d'identifier et de trier les sons qu'il perçoit : il *sépare les différentes sources sonores* et essaie d'extraire celle qui nous intéresse. Cependant, lorsque le bruit environnant devient trop fort, cette faculté n'est plus suffisante. Alors comment expliquer que dans les mêmes conditions sonores nous comprenons très bien ce que dit notre voisin ? La réponse nous vient du monde des sciences cognitives. Nous sommes dotés d'une faculté surprenante, dont pour la plupart d'entre nous ne sommes pas conscients. Lorsque nous parlons à quelqu'un et que nous le regardons, notre cerveau *fusionne ce qu'il entend avec ce qu'il voit*, en particulier le mouvement des articulateurs visibles de la parole, pour nous aider à mieux comprendre : la parole est multimodale.

Formalisée dans le milieu des années 80, par Ch. Jutten, J. Hérault et B. Ans, alors qu'ils travaillaient sur la capacité du cerveau à décoder les informations de vitesse et de position lors d'un mouvement, la séparation de sources est devenue un domaine attractif du traitement du signal. Elle consiste à retrouver des sources inconnues à partir d'observations qui sont des mélanges de celles-ci en exploitant le moins d'information *a priori* possible. La séparation de sources, qui peut aussi être vue comme une généralisation du problème de l'extraction d'un signal utile dans une observation bruitée, a de nombreuses applications tant dans le domaine du traitement des images, des signaux biomédicaux, des télécommunications ou du traitement de la parole. En particulier, le problème qui consiste à extraire un locuteur parmi un mélange donné s'appelle la *cocktail party*. Lorsque la séparation de sources est abordée en faisant l'unique hypothèse d'indépendance mutuelle des sources, elle s'appuie sur l'analyse en composante indépendante (ACI) dont les bases théoriques furent posées au début des années 90 et qui recherche dans les observations à séparer les composantes indépendantes entre elles. Mais ce n'est pas la seule façon de résoudre le problème de séparation de sources : il est possible d'estimer les sources en faisant par exemple l'hypothèse de leur parcimonie dans une certaine représentation, ce qui suppose alors que chaque composante des observations est principalement due à une seule source active. Le problème de séparation de sources s'apparente alors à affecter chaque composante à la bonne source.

Dans cette thèse, on propose une approche originale du problème de l'extraction d'un locuteur dans un mélange de plusieurs sources. Cette approche consiste à utiliser l'information visuelle relative à ce locuteur. Cette étude fait suite à un premier travail réalisé par D. Sodoyer dans le cadre de mélanges additifs instantanés. Dans cette thèse, on étudie le cas plus difficile et plus réaliste des mélanges convolutifs dé-

terminés. Comme on le verra dans cette étude, ce cadre nécessite le développement de nouvelles techniques et des algorithmes associés pour faire face à la plus grande complexité du cadre envisagé.

Notre ambition dans ce travail n'est certes pas de vouloir imiter le fonctionnement de notre cerveau, mais de nous inspirer de ses facultés hors du commun pour proposer dans cette thèse de nouveaux algorithmes de séparation de sources capables d'exploiter la cohérence entre ce que nous entendons et ce que nous voyons. Ainsi, notre travail de thèse porte à la fois sur la modélisation de la multimodalité de la parole et sur son utilisation comme une aide à la séparation de sources de parole.

Organisation du manuscrit

Ce manuscrit est composé de trois parties. La première partie est pour nous l'occasion de faire un bref état de l'art des deux domaines abordés dans cette thèse. La seconde partie est consacrée à la modélisation de la bi-modalité de la parole en vue de son application pour l'extraction de source de parole audiovisuelle qui est abordée dans la troisième partie.

Plus précisément, la première partie est composée de deux chapitres. Le premier chapitre est consacré à la multimodalité de la parole. Nous rappelons que la parole n'est pas qu'auditive : elle est aussi visuelle en ce sens qu'il existe une forte cohérence entre le son prononcé et les mouvements des articulateurs visibles, en particulier celui des lèvres. Nous finissons ce chapitre par un bref aperçu de l'utilisation de la bi-modalité de la parole dans les techniques de traitement du signal appliqué à la parole (par exemple débruitage, compression).

Le deuxième chapitre est consacré à l'introduction à la séparation de sources. Nous abordons dans un premier temps la séparation de sources dans le cas des mélanges linéaires instantanés en présentant quelques-uns des principaux algorithmes fondés soit sur une mesure directe de l'indépendance, soit sur une approximation de l'indépendance par les statistiques d'ordre supérieur. Ensuite, nous présentons la séparation de sources dans des mélanges convolutifs soit dans le domaine temporel soit dans le domaine fréquentiel.

La deuxième partie de ce manuscrit porte sur la modélisation de la multimodalité de la parole. Dans le chapitre 3, nous proposons un modèle statistique audio qui décrit efficacement un son unique de la parole et que nous appelons modèle à loi LogRayleigh. Ce noyau sert de base pour construire un modèle audio multi-noyaux capable de modéliser la parole continue. Finalement, nous étendons ce modèle purement audio à un modèle audiovisuel liant efficacement des paramètres spectraux auditifs à la forme des lèvres du locuteur pour de la parole continue. Une série d'expérimentations montre que notre modélisation multi-LogRayleigh est plus efficace qu'un modèle général à base de noyaux gaussiens pour caractériser les coefficients audio.

Dans le chapitre 4, nous utilisons le modèle audiovisuel du chapitre précédent comme base d'un nouveau détecteur d'activité vocale audiovisuel statistique. Ensuite, nous introduisons la notion de détection de silence (*i.e.* non activité vocale) visuelle reposant sur l'hypothèse du mouvement des lèvres pendant la parole : pendant que nous parlons, nos lèvres bougent tandis que lorsque nous sommes silencieux, elles bougent nettement moins.

La dernière partie de ce manuscrit est consacrée à l'utilisation de la modélisation de la bi-modalité de la parole pour extraire une source particulière de parole de mélanges de type convolutifs. Dans le chapitre 5, nous exploitons tout d'abord le modèle audiovisuel du chapitre 3 pour résoudre le problème des permutations, rencontrées à chaque fréquence, inhérent à tout système de séparation fréquentielle fondée sur l'indépendance. Ensuite, la détection des moments de silence par la modalité visuelle seule permet de résoudre ce même problème des permutations grâce à un algorithme plus simple.

Dans le dernier chapitre de notre étude, nous proposons une nouvelle méthode d'ordre deux pour l'extraction d'une source de parole fondée sur la détection des moments de silence par la modalité visuelle. Pendant les moments de silence détectés, il est possible de déterminer dans les mélanges la direction de la source absente permettant ainsi d'extraire cette source en dehors de ces moments de silence.

Pour terminer, trois annexes complètent ce manuscrit : la première fournit quelques éléments concernant les statistiques d'ordre supérieur. La seconde détaille les calculs de l'étude de la loi LogRayleigh que nous proposons au chapitre 3. Dans la troisième, nous décrivons l'algorithme EM de façon à obtenir les équations d'apprentissage des différents modèles statistiques utilisés.

Première partie

État de l'art

Chapitre 1

Parole audiovisuelle

Affirmer que la parole ne serait pas qu'auditive mais aussi visuelle peut sembler curieux. Pour illustrer ce phénomène, considérons une situation que le lecteur aura certainement déjà vécu. Vous êtes dans un environnement bruyant (une gare par exemple) et vous discutez avec un ami. Il est très probable qu'instinctivement vous regardiez attentivement votre interlocuteur pour mieux comprendre ce qu'il vous raconte. Sans vous en rendre compte, vous êtes en train de lire sur ses lèvres pour vous aider à entendre ce qu'il vous dit. Pour résumer cette situation, nous pouvons reprendre la formule de Bernstein et Benoît [23] :

"Pour percevoir la parole, plusieurs sens valent mieux qu'un."

Dans ce chapitre, nous allons brièvement introduire la notion de bimodalité de la parole. Nous verrons ensuite quel type d'information visuelle est utile avant de montrer que la parole audiovisuelle est à la fois redondante et complémentaire. Enfin, nous présenterons succinctement quelques-unes des applications possibles de la bimodalité de la parole en traitement du signal.

1.1 La parole : un mélange audiovisuel

Pour comprendre que la parole n'est pas qu'auditive, mais que l'information visuelle joue également un grand rôle, nous allons nous intéresser aux personnes malentendantes ou sourdes, chez qui l'aptitude à entendre est réduite ou nulle. Tout le monde sait qu'elles peuvent, en partie, lire sur les lèvres pour les aider à comprendre une discussion. On pourrait penser qu'elles ont développé cette aptitude pour compenser leur défaut d'audition, mais ce ne sont pourtant pas les seules personnes capables de lecture labiale : la grande majorité des personnes voyantes a développé cette faculté de façon instinctive comme l'ont montré les travaux de Sumby et Pollack [125] ou Erber [52]. Ces études, suivies par d'autres comme par exemple [20, 113], ont montré le gain apporté par la vision du locuteur à l'audition de celui-ci pour l'intelligibilité vis-à-vis de l'audition seule. Par exemple, les travaux d'Erber [52] et de Benoît [20] montrent que le taux de reconnaissance correcte de la parole audiovisuelle est supérieure à celui de la parole audio seule (*cf.* figure 1.1). Cette supériorité est d'autant plus grande que le signal acoustique est bruité (*i.e.* pour des rapports signaux sur bruit (RSB) petits). Quand le RSB devient très faible (*i.e.* que le signal audio n'est plus audible) alors les performances de reconnaissance

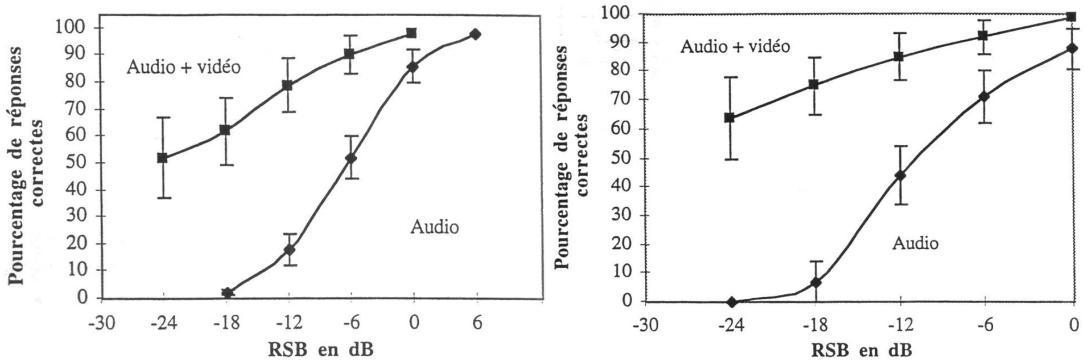


FIG. 1.1 – Influence de la vision dans la reconnaissance de la parole. Taux de reconnaissance correcte auditive et audiovisuelle de la parole acoustiquement bruitée : corpus de 250 mots en anglais (figure 1.1(a)) [52] et 18 logatomes en français (figure 1.1(b)) [20].

correcte des sons tendent vers une valeur correspondant à celles de la lecture labiale seule.

De plus, pour illustrer encore une fois l'influence de la vision du visage d'un locuteur sur ce que nous entendons, intéressons-nous à l'effet McGurk [87]. Cette illusion audiovisuelle consiste à superposer un stimulus [ba] audio à un stimulus [ga] visuel. Dans ces conditions, le conflit entre l'audio et la vidéo aboutit à la perception d'un [da] (l'intensité de l'effet dépendant tout de même du sujet). Bien que cette situation tende à montrer que l'information visuelle influe sur la perception auditive, elle peut n'avoir aucun effet comme dans le cas de film doublé : dans une telle situation la différence entre l'audio et la vidéo est telle, que le spectateur ne cherche plus, même instinctivement, à intégrer les deux modalités.

Finalement, la vision du visage du locuteur permet non seulement de mieux comprendre mais aussi de mieux détecter la parole dans le bruit [61, 79, 22] : le seuil d'audition est abaissé lorsque les sujets voient le visage du locuteur. En effet, la vision renforce les indices acoustiques pertinents et donne l'impression de mieux entendre la personne qui parle. C'est cette idée que *la modalité visuelle de la parole peut être utile pour mieux traiter le signal audio* que nous allons exploiter dans notre étude.

1.2 Information vidéo utile

Maintenant que nous savons que la vision du visage du locuteur influe et peut aider notre audition, demandons-nous quelle partie de l'information visuelle exploite-t-on vraiment ?

Une première idée intuitive est de dire que seules les lèvres sont utiles. Mais ceci est démenti par l'étude de Benoît *et al.* [21]. Ils montrent, par une étude comparative de l'intelligibilité, que les lèvres du locuteur contiennent environ les deux tiers de l'information visuelle véhiculée par tout le visage (*cf.* figure 1.2). Il est possible d'en

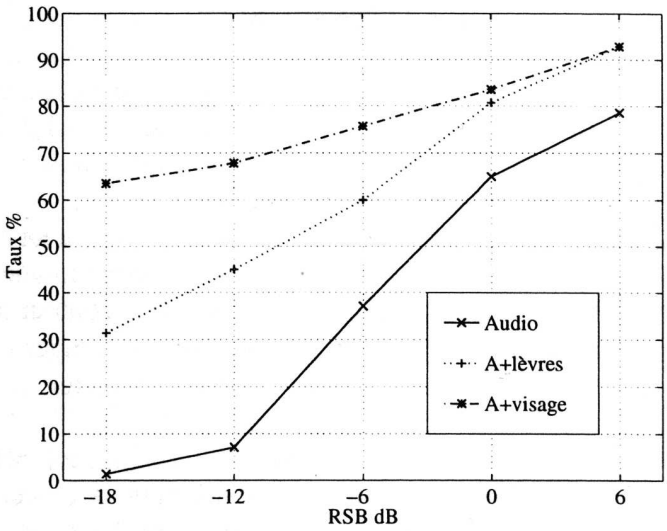


FIG. 1.2 – Étude comparative de l'information visuelle. Taux de reconnaissance correcte en présence de différents stimuli visuels [21].

conclure que l'information visuelle n'est pas seulement contenue dans le mouvement des lèvres mais que d'autres informations sont utiles pour la compréhension de la parole.

Ainsi, pour produire les sons, de nombreuses parties de notre conduit vocal, visibles (par exemple la forme des lèvres ou la position de la mâchoire) et non visibles (par exemple la position de la langue), entrent en œuvre et sont contrôlées. Il est possible de distinguer visuellement un [i] d'un [a] ou bien un [i] d'un [y] alors qu'il est impossible de faire la différence visuelle entre un [y] et un [u]. En effet, dans les deux premiers exemples, la forme des lèvres est différente pour chacun des sons, différence de position de la mâchoire et différence d'ouverture et de protrusion, alors que pour la dernière alternative, seule change la position de la langue ici invisible, la forme des lèvres restant quant à elle identique. Tout comme les phonèmes sont des sons discernables acoustiquement, les visèmes ont été définis comme des formes visuelles discernables [123].

En considérant les résultats précédents, nous pouvons conclure que les lèvres du locuteur véhiculent la majeure partie utile pour la parole de l'information visuelle. Ainsi dans notre étude, les paramètres vidéo que nous exploiterons seront des paramètres relatifs à la forme des lèvres.

1.3 Redondance et complémentarité de la parole audiovisuelle

Nous venons de voir qu'une grande partie de l'information visuelle que nous utilisons pour la parole audiovisuelle est contenue dans les lèvres du locuteur. Intéressons-

nous maintenant aux relations existant entre ces paramètres vidéo et des paramètres audio.

1.3.1 Redondance

Intuitivement, nous pouvons prédire qu'il doit y avoir une cohérence entre les mouvements du visage du locuteur, et plus particulièrement ceux des ses lèvres, et le son émis. En effet, ces deux phénomènes sont produits par un seul et même système : les articulateurs. Ainsi, [137] a pour but de montrer qu'il existe une relation entre la production et la perception multimodale de la parole. Pour cela, les auteurs étudient les relations linéaires qui peuvent exister entre le visage du locuteur (18 marqueurs placés sur la face), son conduit vocal (4 capteurs placés sur la langue) et le son produit (coefficients LSP *line spectrum pairs* et la puissance du signal). Leurs études montrent qu'une grande partie de la variance totale de la face d'un locuteur peut être prédite linéairement à partir de son conduit vocal ($\sim 90\%$), mais que la prédiction inverse est aussi vérifiée ($\sim 80\%$). De même, ils montrent qu'une partie ($\sim 75\%$) de l'enveloppe spectrale des sons produits peut être prédite linéairement à partir du visage du locuteur. Cependant, ces résultats sont à interpréter avec précaution. En effet, [10] montre que si l'on utilise la seule forme des lèvres comme information visuelle, les résultats de la prédiction linéaire de l'enveloppe spectrale du son produit chutent ($\sim 50\%$) mais qu'ils peuvent être améliorés ($\sim 60\%$) en choisissant une prédiction non-linéaire. Tous ces travaux montrent cependant qu'il existe une certaine cohérence entre la forme des lèvres et les sons produits.

1.3.2 Complémentarité

La cohérence entre le son et l'image n'est pas totale. En effet, comme nous allons le voir, il y a également une certaine complémentarité entre ces deux modalités. Sans l'avoir mentionnée explicitement, nous avons déjà abordé cette notion de complémentarité. En effet, au paragraphe 1.1 nous avons vu que la multimodalité de la parole permettait d'améliorer les performances de reconnaissance par rapport à la seule modalité auditive. Cette propriété est également illustrée dans [126] grâce aux arbres de confusions (*cf.* figures 1.3 et 1.4). Cela consiste à présenter des stimuli à des sujets adultes et bien entendants puis de classifier les confusions faites entre ces stimuli en fonction du niveau de bruit environnant. L'analyse de ces arbres de confusion montre que deux consonnes voisines auditivement, [k] et [p] ou [m] et [n] par exemple, sont bien distinctes visuellement. Cette complémentarité pour les consonnes a été montrée ensuite pour les voyelles [113]. La figure 1.5 traduit géométriquement la distance perceptive auditive et visuelle entre les voyelles du français. Ces schémas montrent que des voyelles proches auditivement sont éloignées visuellement.

Finalement, la redondance et la complémentarité audiovisuelles de la parole ne sont que partielles et les relations entre les paramètres vidéo et audio ne peuvent pas être envisagées de façon linéaire car complexes.

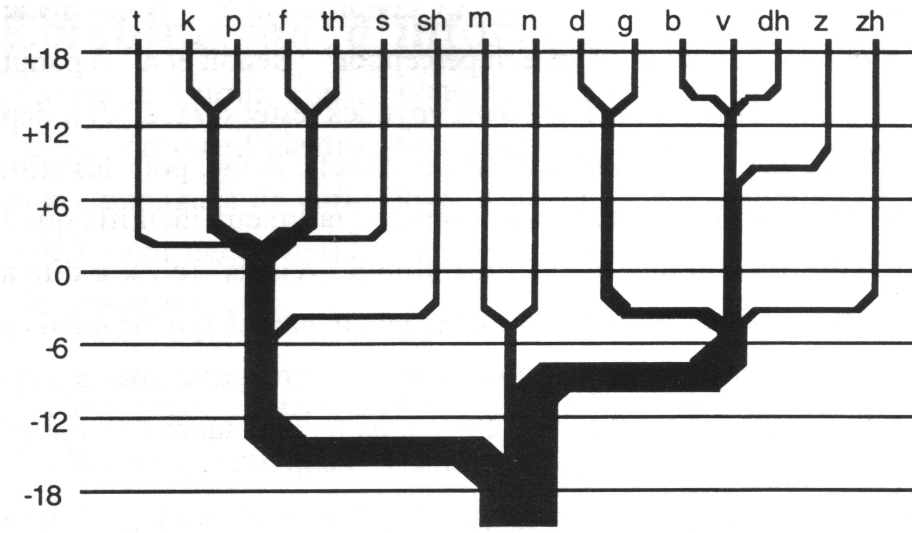


FIG. 1.3 – Arbres de confusion auditive des consonnes en fonction du RSB (en dB) [126].

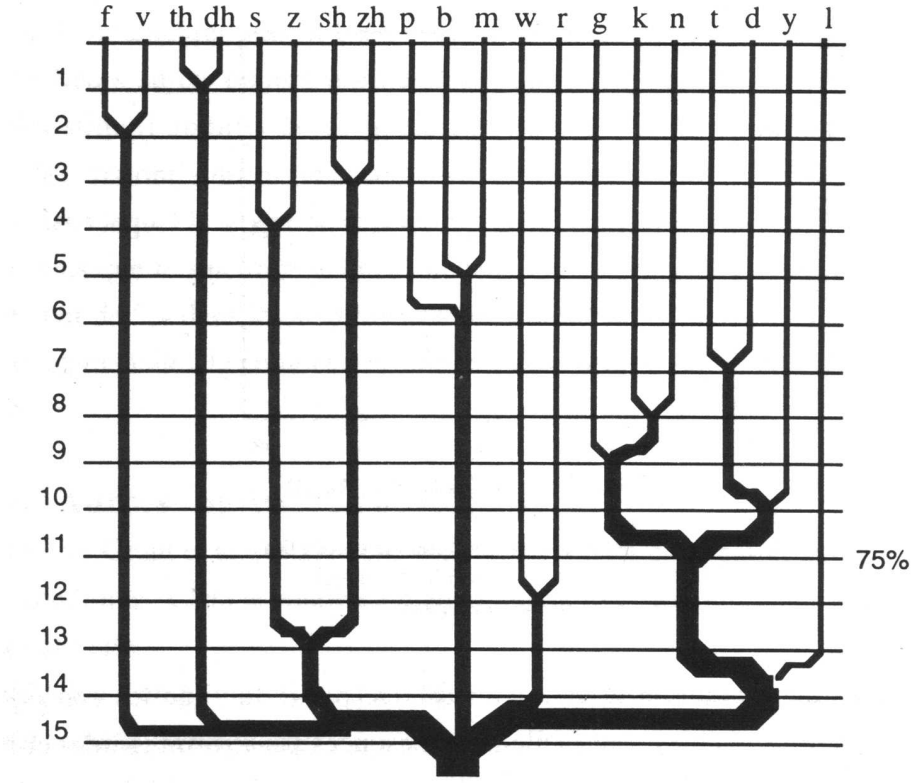
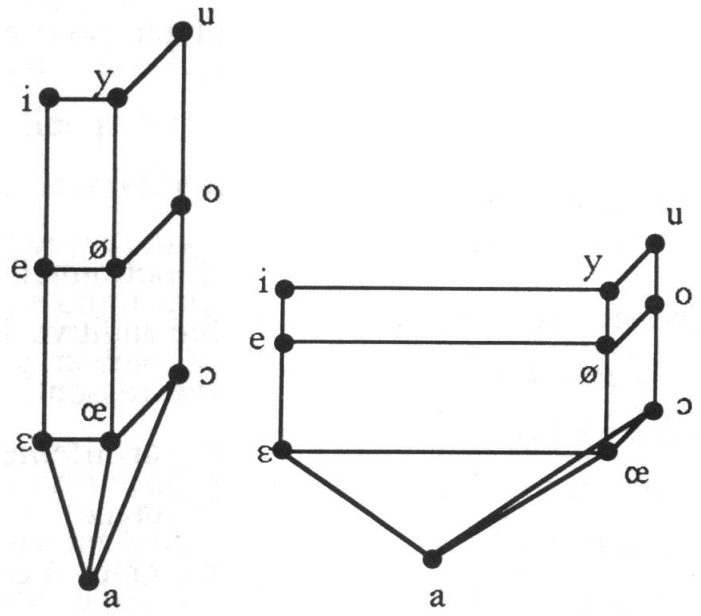


FIG. 1.4 – Arbres de confusion visuelle des consonnes. L'échelle verticale correspond au niveau de regroupement [126].



(a) (b)

FIG. 1.5 – Schémas de la géométrie auditive (figure 1.5(a)) et visuelle (figure 1.5(b)) [113].

1.4 Bi-modalité de la parole en traitement du signal

Comme nous venons de le voir, la parole est (au moins) bimodale car audiovisuelle. Cette propriété intrinsèque de la parole a été mise à profit dans des systèmes de traitement du signal de façon à en améliorer les performances. Nous décrivons brièvement les exemples de la reconnaissance automatique de la parole, du débruitage de signaux ou de la compression.

1.4.1 Reconnaissance automatique de la parole

Une première application en traitement du signal à avoir recours à la bimodalité de la parole est celle de la reconnaissance automatique de la parole (RAP). En effet, comme nous l'avons vu au paragraphe 1.1, l'emploi de la modalité visuelle permet d'augmenter les scores de reconnaissance pour les individus. Il a été naturel d'essayer de reproduire cette amélioration pour les procédés automatiques. Ainsi, de nombreux algorithmes ont été proposés depuis les premiers travaux de Petajan en 1984 [94] (*cf.* [104] pour une revue de la littérature). Ils ont tous le même schéma de principe (*cf.* figure 1.6) : extraction des paramètres audio et vidéo, intégration audiovisuelle de ces données puis le système de reconnaissance à proprement parler. La façon de procéder pour l'intégration audiovisuelle diffère d'un algorithme à l'autre. Ainsi, certains utilisent une fusion des paramètres audio et vidéo utilisés [129], tandis que d'autres vont plutôt intégrer les décisions obtenues par deux systèmes unimodaux (audio et vidéo séparément) [49] pour reconnaître la parole audiovisuelle.

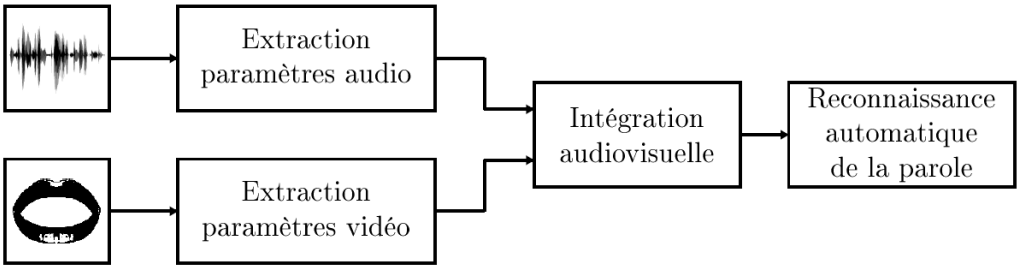


FIG. 1.6 – Schéma de principe de la reconnaissance automatique de la parole.

1.4.2 Débruitage et séparation de sources audiovisuelles

Une autre application possible est celle du débruitage ou réhaussement de la parole. Elle consiste à estimer le signal de parole originel $s(t)$ à partir d'une observation bruitée $x(t)$ de celui-ci : $x(t) = s(t) + b(t)$, où $b(t)$ est le bruit. Quand plusieurs capteurs sont disponibles, le problème de réhaussement de parole peut s'apparenter à celui de la séparation de sources (*cf.* chapitre 2 pour l'étude de la séparation de sources). Cette application de la multimodalité de la parole occupant une place particulière dans notre étude, nous la développerons plus en détails au paragraphe 2.4.

Puisqu'il existe une cohérence entre le son et l'image d'une part et que, d'autre part, les caractéristiques sonores peuvent être partiellement prédites à partir de l'image, ces systèmes de débruitage vont exploiter cette opportunité pour estimer des filtres de réhaussement [59] par une prédiction linéaire des paramètres audio à partir des paramètres vidéo :

$$\mathbf{a}(t) = M \begin{bmatrix} 1 \\ \mathbf{v}(t) \end{bmatrix} \tag{1.1}$$

où $\mathbf{a}(t)$ et $\mathbf{v}(t)$ sont les vecteurs des paramètres audio et vidéo respectivement et M est la matrice de prédiction. Cette idée fut ensuite reprise en utilisant des outils plus sophistiqués en pré-traitement d'un système de reconnaissance de la parole [45, 60]. Récemment, ce principe fut étendu à celui de la séparation de sources de parole audiovisuelle [118, 134]. L'information visuelle peut alors être utilisée au travers d'un modèle statistique audiovisuel $p_{AV}(\mathbf{a}(t), \mathbf{v}(t))$ reliant des paramètres audio $\mathbf{a}(t)$ à des paramètres vidéo $\mathbf{v}(t)$ [118]. Le principe consiste alors à retrouver en sortie du système de séparation le son le plus cohérent avec la vidéo en maximisant cette probabilité audiovisuelle. Ou alors, l'information visuelle est utilisée pour contraindre le problème de séparation [134].

1.4.3 Compression audiovisuelle

La dernière application dont nous parlerons est celle du codage de la parole audiovisuelle : elle consiste à coder conjointement les signaux audio et vidéo [58] alors que plus classiquement les deux modalités de la parole le sont séparément. Le but est de compresser de façon plus efficace les signaux pour améliorer les débits de

transmission en visiophonie par exemple. Cette application exploite la redondance de la parole de façon à ne coder qu'une seule fois une information présente à la fois dans l'audio et la vidéo. Cette application est un peu particulière car elle peut être vue comme faisant le contraire des autres : la redondance ou la complémentarité n'est pas vue ici comme un atout mais comme une nuisance que l'on cherche sinon à supprimer tout du moins à minimiser.

1.5 Conclusion

Ce chapitre nous a permis d'avoir un rapide aperçu de la notion de multimodalité de la parole depuis la perception jusqu'à son intégration dans des applications du traitement du signal qui exploitent redondance et complémentarité entre les modalités auditive et visuelle. Nous pouvons donc conclure ce chapitre en disant que la parole n'est pas qu'auditive et que la modalité visuelle nous permettra de mieux traiter les signaux acoustiques.

Chapitre 2

Séparation aveugle de sources

La séparation de source est un domaine relativement récent du traitement du signal. Introduite dans le milieu des années 80 par Ans, Hérault et Jutten [7, 66] alors qu'ils travaillaient sur un problème biologique, la séparation de source est très vite devenue un domaine attractif du traitement du signal (*c.f.* [77] pour des considérations historiques). Le problème consiste à retrouver des signaux utiles (par exemple signaux de parole ou des signaux émis par des téléphones portables), aussi appelés *sources*, à partir de mélanges, aussi appelés *observations*, de ceux-ci. Généralement, les observations sont des signaux obtenus à partir d'un ensemble de capteurs (microphones ou antennes par exemple). Un cas typique est celui de la *cocktail party* où les sources sont des locuteurs et les observations les signaux enregistrés par des microphones (*c.f.* figure 2.1). Dans un contexte *aveugle*[1], aucune connaissance *a priori* n'est disponible ni sur les sources, ni sur le *processus de mélange* (*i.e.* le contexte des observations), cette situation est alors appelée *séparation aveugle de source* (SAS). Pour résoudre ce problème, une solution possible consiste à ne faire qu'une seule hypothèse fondamentale : *l'indépendance statistique mutuelle des sources.*

Le succès de la séparation de sources s'explique par le peu d'information *a priori* nécessaire pour résoudre ce problème et par le vaste champ d'applications possibles par exemple le traitement de signaux biomédicaux (avec entre autres l'extraction de signaux électrocardiogrammes d'un fœtus [43, 140], ou la suppression des artefacts pour l'analyse des signaux électroencéphalogrammes du cerveau [78]), de signaux vibratoires de machines tournantes [25], de signaux pour la surveillance d'aéroport [33], de signaux de télécommunication [131], de signaux acoustiques [130, 6] pour ne citer que celles-là (*c.f.* [69, 5] pour d'autres applications).

Dans ce chapitre, nous présentons de façon formelle le problème de la séparation de sources avant de voir les conditions de séparabilité et les indéterminations intrinsèques au problème. Nous détaillerons ensuite deux situations typiques de mélanges, les mélanges instantanés et convolutifs, en présentant pour chacune des situations les principes de séparation.

[1]Sans aucune information *a priori*, ni sur les sources ni sur le processus de mélange, ce problème n'admet pas de solution.

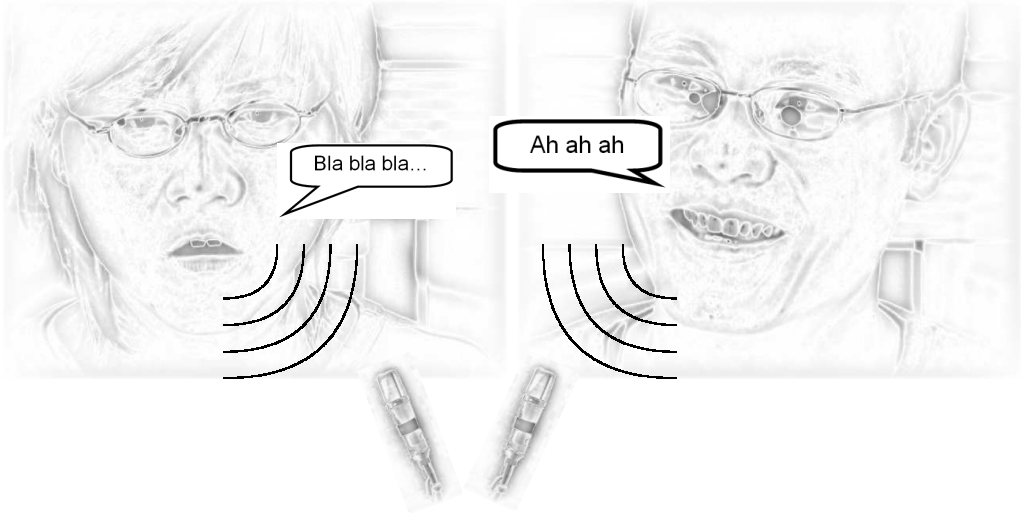


FIG. 2.1 – Exemple de la *cocktail party* avec deux sources et deux capteurs.

2.1 Présentation générale de la séparation de sources

Comme nous l'avons brièvement introduit ci-dessus, le problème de la séparation aveugle de source consiste à retrouver, avec le moins de connaissance *a priori*, des signaux utiles qui ont été mélangés. Formalisons maintenant cette idée.

2.1.1 Formulation mathématique

Supposons que nous ayons à notre disposition N_o observations, notées $\mathbf{x}(t) = [x_1(t), \cdots, x_{N_o}(t)]^T$, de N_s sources, notées $\mathbf{s}(t) = [s_1(t), \cdots, s_{N_s}(t)]^T$, obtenues à partir d'une fonction de mélange $\mathcal{H}(\cdot)$

$$\mathbf{x}(t) = \mathcal{H}(\mathbf{s}(t)). \tag{2.1}$$

Dans le cas général, $\mathcal{H}(\cdot)$, qui est une application de $\mathcal{E}_{N_s}$, espace des sources de dimension N_s, dans $\mathcal{E}_{N_o}$, espace des observations de dimension N_o, peut être non-linéaire et à mémoire ($\mathcal{H} : \mathcal{E}_{N_s} \rightarrow \mathcal{E}_{N_o}$). Diverses situations peuvent intervenir suivant le nombre N_o d'observations relativement au nombre N_s de sources :

- moins d'observations que de sources ($N_o < N_s$), on parle alors de *mélange sous-déterminé*,
- autant d'observations que de sources ($N_o = N_s$), le mélange est dit *déterminé*,
- plus d'observations que de sources ($N_o > N_s$), le mélange est qualifié de *sur-déterminé*.

Ces trois cas supposent des conditions sur $\mathcal{H}(\cdot)$. De plus si le processus de mélange $\mathcal{H}(\cdot)$ est linéaire, nous le qualifierons assez naturellement de *mélange linéaire* et de *mélange non linéaire* dans le cas contraire.

Le but de la SAS étant de retrouver les sources à partir uniquement des observations $\mathbf{x}(t)$ et en exploitant l'hypothèse d'indépendance mutuelle des sources[2], il est

[2]Éventuellement d'autres informations *a priori* sur les sources ou le processus de mélange pourront être introduites, au cas par cas, suivant les connaissances *a priori* dont nous disposons.

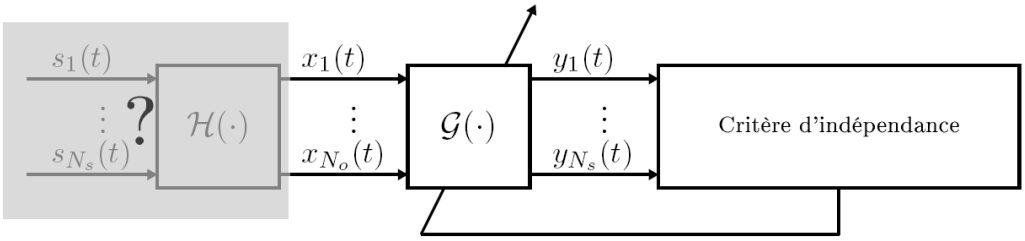


FIG. 2.2 – Principe de la séparation aveugle de source.

alors nécessaire de construire une *fonction de séparation* $\mathcal{G}(\cdot)$ telle que chacune de ses sorties

$$\mathbf{y}(t) = \mathcal{G}(\mathbf{x}(t)) \tag{2.2}$$

ne dépende que d'une source et une seule. Dans le cas général, $\mathcal{G}(\cdot)$, qui est une application d'un espace de dimension N_o dans un espace de dimension N_s, peut elle aussi être non linéaire et à mémoire.

Puisque la seule hypothèse dont nous disposons est l'indépendance mutuelle des sources, il est naturel d'essayer de chercher une fonction de séparation $\mathcal{G}(\cdot)$ tel que son vecteur de sortie $\mathbf{y}(t)$ ait des composantes les plus indépendantes possible. La figure 2.2 montre le schéma synoptique général de la séparation aveugle de sources.

2.1.2 Séparabilité et indéterminations

Séparabilité

La question primordiale est maintenant celle de la *séparabilité* des mélanges (*i.e.* l'existence d'une solution) : "l'indépendance des composantes de $\mathbf{y}(t)$ implique-t-elle nécessairement la séparation des sources ?" En d'autres termes, l'indépendance des composantes de $\mathbf{y}(t)$ implique-t-elle que chacune des sorties de la fonction de séparation ne dépend que d'une et une seule source.

Autrement dit, existe-t-il des transformations $\mathcal{G}(\cdot)$ qui sont mélangeantes, c'est-à-dire telle que $(\mathcal{G} \circ \mathcal{H})(\cdot)$ soit à Jacobien non diagonal, et qui préservent l'indépendance ? Malheureusement, la réponse à cette question est généralement oui sauf dans certains cas particuliers sur lesquels nous reviendrons ultérieurement : l'indépendance n'est pas suffisante pour garantir la séparation des sources. Nous illustrons, ci-dessous, ceci sur un exemple simple, mais Darmois [41] propose une méthode simple de construction de telles transformations.

Considérons deux sources s_1 et s_2, indépendantes et identiquement distribuées (iid) normalement telles que $s_1 \overset{iid}{\sim} \mathcal{N}(0,1)$ et $s_2 \overset{iid}{\sim} \mathcal{N}(0,1)$. Supposons maintenant que les observations $\mathbf{x} = [x_1, x_2]^T$ vérifient

$$\begin{aligned} x_1 &= \cos(\theta)\, s_1 - \sin(\theta)\, s_2 \\ x_2 &= \sin(\theta)\, s_1 + \cos(\theta)\, s_2. \end{aligned}$$

Ces observations sont gaussiennes, puisque sommes de variables aléatoires gaussiennes indépendantes, et leur matrice de covariance $C_{\mathbf{xx}}$ vérifie

$$C_{\mathbf{xx}} \triangleq E\left[\mathbf{x}\mathbf{x}^T\right] = I_2,$$

où I_n est la matrice identité de taille $(n \times n)$. Ainsi, les observations $\mathbf{x}$ sont décorrélées et donc indépendantes puisque gaussiennes. Dans ce cas, choisir $\mathcal{G}(\cdot) = I_2$ permet bien d'obtenir des sorties $\mathbf{y}$ mutuellement indépendantes ($\mathbf{y} = \mathbf{x}$). Or chacune des composantes de $\mathbf{y}$ ne dépend que d'une seule source que lorsque θ est égal à zéro modulo $\pi/2$: $\theta \equiv 0[\pi/2]$. Donc, dans tous les autres cas, on obtient des sorties mutuellement indépendantes sans pour autant satisfaire la séparation des sources. Ceci illustre bien que, d'une manière générale, l'indépendance mutuelles des composantes de $\mathbf{y}(t)$ n'implique pas nécessairement la séparation des sources.

Ainsi, nous ne pouvons donner de résultats généraux sur la séparabilité des mélanges : il nous faudra donc faire une étude au cas par cas.

Indéterminations

Admettons cependant que le mélange que nous étudions soit séparable, l'existence d'une solution (*i.e.* la séparabilité) assure-t-elle son unicité ? Pour cela supposons que $\mathbf{y}(t)$ soit un vecteur solution. Il a été obtenu uniquement grâce à un critère d'indépendance de ses composantes, or cette indépendance n'impose aucune contrainte sur l'ordre de celles-ci : si $\mathbf{y}(t)$ est un vecteur solution alors $\mathbf{y}'(t) = \Pi\,\mathbf{y}(t)$, où Π est une matrice de permutation, est aussi un vecteur solution car ayant ses composantes indépendantes. Nous venons de mettre en évidence la première indétermination celle de la *permutation* : les sources ne pourront être estimées qu'à une permutation globale près.

De plus le critère d'indépendance des composantes du vecteur solution $\mathbf{y}(t)$ n'implique aucune contrainte sur une éventuelle déformation de celles-ci : si $\mathbf{y}(t)$ est un vecteur solution alors $\mathbf{y}'(t) = \Lambda(\mathbf{y}(t))$, où $\Lambda(\cdot)$ est une matrice diagonale de fonctions (linéaires ou non), est aussi un vecteur solution. Nous venons de mettre en évidence la seconde indétermination celle du *facteur d'échelle* : les sources ne pourront être estimées qu'à une distorsion près.

Définition 2.1 (Egalité séparante)
Nous dirons que le vecteur $\mathbf{x}(t)$ *est* égal au sens séparant *au vecteur* $\mathbf{y}(t)$*, ce que nous notons* $\mathbf{x}(t) \cong \mathbf{y}(t)$*, si et seulement si* $\mathbf{x}(t)$ *est égal à* $\mathbf{y}(t)$ *à une permutation* Π *et une distorsion diagonale* $\Lambda(\cdot)$ *près :*

$$\mathbf{x}(t) \cong \mathbf{y}(t) \quad \overset{\triangle}{\Longleftrightarrow} \quad \exists\,\Pi, \Lambda(\cdot) \;/\; \mathbf{x}(t) = \Pi\,\Lambda(\mathbf{y}(t)). \tag{2.3}$$

Définition 2.2 (Fonction séparante)
Nous appellerons fonction séparante *toute fonction de séparation* $\mathcal{G}(\cdot)$ *tel que ses sorties* $\mathbf{y}(t) = \mathcal{G}(\mathbf{x}(t))$*, où* $\mathbf{x}(t)$ *sont des observations de sources* $\mathbf{s}(t)$*, soient égales au sens séparant aux sources* $\mathbf{s}(t)$ *:* $\mathbf{y}(t) \cong \mathbf{s}(t)$*. Nous dirons alors, par abus de langage, que*

$$(\mathcal{G} \circ \mathcal{H})(\cdot) = \Pi\,\Lambda(\cdot) \cong I_{N_s}. \tag{2.4}$$

Nous pouvons donc résumer la séparabilité et les deux indéterminations, permutation et facteur d'échelle, de la façon suivante :

> *Si une solution au problème de la séparation de sources existe alors elle vérifie*
>
> $$\hat{\mathbf{s}}(t) = \Pi\,\Lambda(\mathbf{s}(t)) \cong \mathbf{s}(t). \tag{2.5}$$

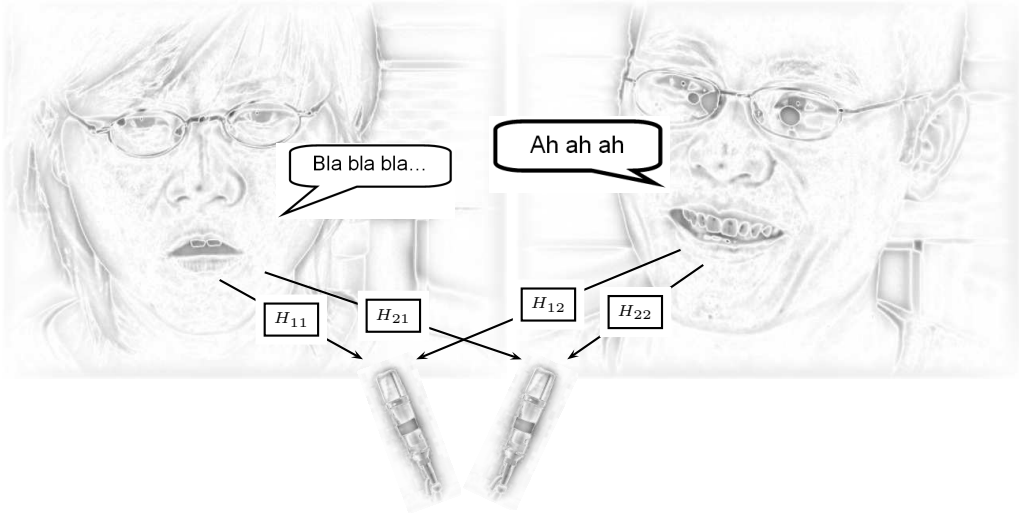


FIG. 2.3 – *Cocktail party* dans le cadre de mélanges linéaires instantanés : les canaux de transmission entre les sources et les capteurs sont modélisés par de simples coefficients $H_{i,j}$.

Ceci signifie concrètement que, sous la condition d'existence d'une solution et sans autre information *a priori* que l'indépendance mutuelle des sources, il n'est possible au mieux de les estimer qu'à une permutation globale près Π et à une distorsion diagonale près $\Lambda(\cdot)$.

La séparation aveugle de sources consiste donc à chercher une fonction séparante $\mathcal{G}(\cdot)$ en s'appuyant uniquement sur l'hypothèse fondatrice de la SAS : l'indépendance statistique mutuelle des sources.

2.2 Mélanges linéaires instantanés

Considérons maintenant le cas particulier des mélanges *linéaires instantanés* dans lequel la fonction de mélange $\mathcal{H}(\cdot)$ est supposée *linéaire* et *sans mémoire* (*cf* Figure 2.3). Les N_s sources, statistiquement indépendantes, sont donc supposées être mélangées de façon linéaire et instantanée : chacune des N_o observations peut ainsi s'exprimer par

$$x_i(t) = \sum_{j=1}^{N_s} H_{i,j}\, s_j(t), \quad \forall i \in \{1, \cdots, N_o\} \tag{2.6}$$

où les $H_{i,j}$ sont des constantes de mélange inconnues. Il est possible de réécrire ce modèle sous forme matricielle

$$\mathbf{x}(t) = H\, \mathbf{s}(t) \tag{2.7}$$

en faisant apparaître la *matrice de mélange* H, de dimension $(N_o \times N_s)$, qui a pour $(i,j)^{\text{ème}}$ élément la constante $H_{i,j}$. La séparation de sources consiste alors à estimer une *matrice de séparation* G, de dimension $N_s \times N_o$, telle que ses sorties

$$\mathbf{y}(t) = G\, \mathbf{x}(t) \cong \mathbf{s}(t) \tag{2.8}$$

soient des estimées des sources originales $\mathbf{s}(t)$. En d'autres termes, G, dont les coefficients sont estimés grâce à l'hypothèse d'indépendance des sources, doit être une matrice séparante.

2.2.1 Séparabilité et indéterminations

Le cas des mélanges linéaires instantanés (2.7) peut être vu comme la résolution d'un système linéaire de N_o équations (celles définissant les observations) à N_s inconnues (les sources). Dans le contexte aveugle, les coefficients de ce système d'équations (ici la matrice de mélanges H) sont également inconnus. Ainsi, les mélanges sous-déterminés, sans autre information *a priori*, ne sont pas séparables puiqu'ils présentent plus d'inconnues (les sources) que d'équations (les observations). D'autre part, les mélanges déterminés et sur-déterminés admettent *a priori* une solution telle que

$$G\,H \cong I_{N_s} \tag{2.9}$$

si H est de rang plein. Cependant, la seule indépendance statistique mutuelle des composantes de $\mathbf{y}(t)$ défini par (2.8) assure-t-elle la séparation du mélange, *i.e* $G \cong H^{-1}$ ou $G \cong H^{\dagger}$ († est la pseudo-inverse d'une matrice de rang plein) pour respectivement les mélanges déterminés ou sur-déterminés ? Autrement dit, existe-t-il des fonctions G non séparantes qui préservent l'indépendance de leurs sorties ? Comon [38] a prouvé que "si au plus une source est gaussienne, alors l'indépendance conjointe (ou paire par paire) des composantes de $\mathbf{y}(t)$ implique que $G\,H = \Pi\,\Lambda$ où Π est une matrice de permutation et Λ une matrice diagonale". Ce théorème, qui est une conséquence du théorème de Darmois-Skitovich de 1953 [42], revient à dire que si au plus une source est gaussienne alors le mélange déterminé (ou sur-déterminé) est séparable et que les sources seront estimées à une permutation globale Π et un gain Λ près. Notez que dans le cas linéaire instantané, la distorsion $\Lambda(\cdot)$ se résume à une simple matrice diagonale : l'indétermination d'échelle se traduit ici par une indétermination sur la puissance des sources reconstituées.

2.2.2 Principe de séparation

Nous allons maintenant exposer les idées fondamentales utilisées pour effectuer la séparation des mélanges instantanés linéaires déterminés[3] où $N_s = N_o$. Nous supposerons de plus que la matrice de mélange est de rang plein. La séparation de sources se résume alors à estimer une matrice de séparation inversible G de taille $(N_s \times N_s)$.

Indépendance statistique et information mutuelle

Rappelons tout d'abord la définition de l'indépendance statistique. N variables aléatoires $\{Y_i\}_{1\leq i\leq N}$ sont mutuellement indépendantes si et seulement si la densité de probabilité conjointe $p_{Y_1,\cdots,Y_N}[y_1,\cdots,y_N]$ est égale au produit des densités de

[3] Les mélanges sur-déterminés pouvant se ramener à ce cas en réduisant le nombre d'observations au nombre de sources.

probabilités marginales $p_{Y_i}[y_i]$ de chacune des variables aléatoires

$$Y_1, \cdots, Y_N \text{ indépendantes} \quad \overset{\triangle}{\Longleftrightarrow} \quad p_{Y_1,\cdots,Y_N}[y_1, \cdots, y_N] = \prod_{i=1}^{N} p_{Y_i}[y_i]. \tag{2.10}$$

Autrement dit, l'indépendance de variables aléatoires se traduit par le fait que la densité de probabilité conjointe est *séparable* ou *factorisable*. Ainsi, les diverses méthodes de séparation exploitant l'indépendance devront être construites de telle sorte que les sources estimées vérifient (ou au moins approximent) cette propriété.

Néanmoins, l'utilisation directe de la définition de l'indépendance n'est pas aisée puisque faisant intervenir des fonctions multivariées (inconnues). Pour cela, une mesure scalaire de l'indépendance, plus pratique, est la divergence de Kullback-Leibler $KL[\cdot\|\cdot]$ entre deux densités de probabilité $p[\cdot]$ et $q[\cdot]$, définie par

$$KL[p\|q] \triangleq \int p[u] \ \ln\left(\frac{p[u]}{q[u]}\right) \ du. \tag{2.11}$$

On peut montrer que cette divergence est une grandeur positive qui s'annule si et seulement si les densités de probabilités $p[\cdot]$ et $q[\cdot]$ sont égales. Ainsi, l'indépendance des composantes du vecteur aléatoire[4] $\mathbf{y} = [y_1, \cdots, y_N]^T$ peut être mesurée par l'information mutuelle $I[\mathbf{y}]$ [39] définie comme la divergence de Kullback-Leibler entre $p_{\mathbf{y}}[\cdot]$ et $\prod_i p_{y_i}[\cdot]$:

$$I[\mathbf{y}] \triangleq KL\left[p_{\mathbf{y}} \left\| \prod_{i=1}^{N} p_{y_i}\right.\right] = \int p_{\mathbf{y}}[\mathbf{u}] \ \ln\left(\frac{p_{\mathbf{y}}[\mathbf{u}]}{\prod_{i=1}^{N} p_{y_i}[u_i]}\right) \ d\mathbf{u}. \tag{2.12}$$

L'information mutuelle $I[\mathbf{y}]$ peut être exprimée par

$$I[\mathbf{y}] = \sum_{i=1}^{N} H[y_i] - H[\mathbf{y}] \tag{2.13}$$

où $H[y_i]$ et $H[\mathbf{y}]$ sont les entropies de Shannon[5] marginales et conjointe respectivement :

$$\begin{aligned} H[\mathbf{y}] &\triangleq -\int p_{\mathbf{y}}[\mathbf{u}] \ \ln(p_{\mathbf{y}}[\mathbf{u}]) \ d\mathbf{u}, \\ H[y_i] &\triangleq -\int p_{y_i}[u_i] \ \ln(p_{y_i}[u_i]) \ du_i. \end{aligned}$$

Notons que l'entropie de Shannon peut être exprimée à partir de l'espérance du logarithme népérien de la densité de probabilité de la variable aléatoire : $H[\mathbf{y}] = -\mathrm{E}[\ln(p_{\mathbf{y}}[\mathbf{y}])]$. L'information mutuelle $I[\mathbf{y}]$ quantifiant l'indépendance des composantes du vecteur aléatoire $\mathbf{y}$, de nombreux algorithmes de séparation de sources y sont explicitement ou implicitement reliés, comme nous allons le voir.

[4]Dans toute la suite de ce manuscrit et par abus de langage, nous confondrons les notations de la variable aléatoire Y avec sa réalisation y.

[5]La définition de l'entropie de Shannon fait souvent intervenir les logarithmes binaires, en choisissant les logarithmes népériens les deux définitions ne diffèrent que d'un facteur multiplicatif.

De l'analyse en composantes principales à l'analyse en composantes indépendantes

De nombreuses méthodes du traitement du signal se concentrent sur l'utilisation des statistiques d'ordre 2 des signaux considérés, comme par exemple le filtrage de Wiener [92]. Appliquer les statistiques du second ordre dans le cadre de la séparation de sources, revient à *décorréler* les mélanges, c'est-à-dire à estimer des signaux centrés $\mathbf{z} = W\,\mathbf{x}$ tels que leur matrice de covariance $C_{\mathbf{zz}} = \mathrm{E}[\mathbf{z}\mathbf{z}^T]$ soit diagonale. En effet, puisque les sources $\mathbf{s}$ sont supposées indépendantes (donc décorrélées) et sans perte de généralité centrées, alors leur matrice de covariance $C_{\mathbf{ss}} = \mathrm{E}[\mathbf{s}\mathbf{s}^T]$ est diagonale. De plus, chaque élément diagonal représente la puissance moyenne de la source correspondante. Donc pour séparer les sources (*i.e.* rechercher des signaux indépendants), il est nécessaire que les sources estimées $\mathbf{z}$ soient décorrélées.

La *décorrélation*, encore appelée *blanchiment* ou *analyse en composantes principales* (ACP), a pour objectif d'estimer des signaux $\mathbf{z}$ dont la matrice de covariance est diagonale. Cette décorrélation peut être réalisée par la décomposition en valeur propre de la matrice de covariance $C_{\mathbf{xx}}$ des observations ou par la décomposition de Cholesky. En effet, la matrice de covariance $C_{\mathbf{xx}}$, qui est symétrique (ou hermitienne si les signaux sont complexes mais nous ne traiterons ici que le cas des signaux réels), est diagonalisable :

$$\exists\, V \in \mathrm{U}(N_s),\ \exists\, D \in \mathrm{D}(N_s)\ /\ C_{\mathbf{xx}} = V\,D\,V^T, \tag{2.14}$$

où $\mathrm{U}(n)$ est le groupe des matrices unitaires de taille $(n \times n)$ et $\mathrm{D}(n)$ l'ensemble des matrices diagonales de taille $(n \times n)$. Les termes diagonaux de D sont les valeurs propres de la matrice de covariance $C_{\mathbf{xx}}$ et les colonnes de V sont les vecteurs propres associés. Ainsi, choisir une *matrice de blanchiment* (spatial) W telle que

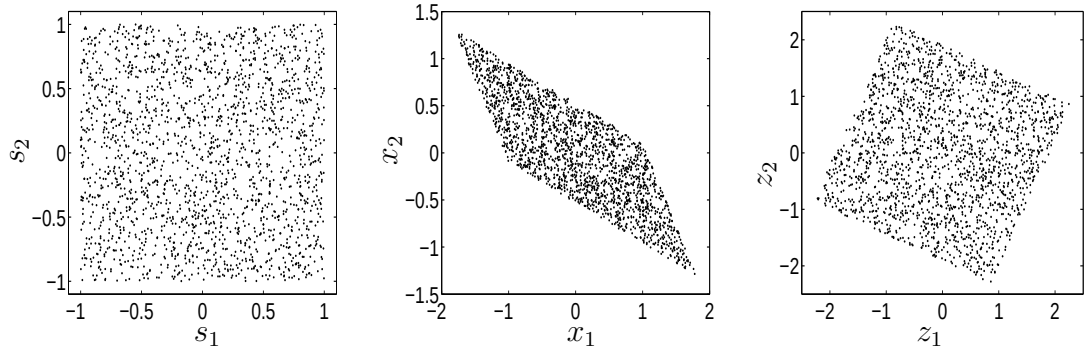
$$W = D^{-\frac{1}{2}}\,V^T \tag{2.15}$$

permet d'effectuer la décorrélation. En imposant de plus, de façon arbitraire, que la matrice de covariance $C_{\mathbf{zz}}$, des *signaux blanchis* définis par $\mathbf{z} = W\,\mathbf{x}$, soit l'identité, nous obtenons :

$$C_{\mathbf{zz}} = \mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right] = W\,C_{\mathbf{xx}}\,W^T = I_{N_s}$$

qui est obtenu en remplaçant W par son expression (2.15) et en utilisant la décomposition en valeurs propres de la matrice de covariance des observations (2.14). Remarquons que le fait d'imposer la puissance moyenne des signaux $\mathbf{z}$ à un, revient à fixer l'indétermination du facteur d'échelle : quelle que soit la matrice de blanchiment W (définie par (2.15) ou par toute matrice obtenue par multiplication à gauche de (2.15) par une matrice diagonale et/ou une matrice de permutation), la normalisation de la puissance des signaux estimés permet d'obtenir toujours la même solution, levant ainsi l'indétermination de gain sans pour autant la résoudre (sauf dans le cas de sources de puissance unité). Les composantes principales sont donc obtenues en projetant les observations $\mathbf{x}$ sur les vecteurs propres de la matrice de covariance $C_{\mathbf{xx}}$ des mélanges fournissant ainsi des signaux décorrélés.

Cependant, bien que la décorrélation soit nécessaire à l'indépendance elle n'en demeure pas moins insuffisante comme illustré à la figure 2.4. Malgré leur décorrélation, les mélanges blanchis $\mathbf{z}$ ne sont pas égaux au sens séparant aux sources. En



(a) Sources (b) Mélanges (c) Mélanges blanchis

FIG. 2.4 – Illustration de l'ACP. Distributions conjointes : de deux sources indépendantes uniformément distribuées entre -1 et 1 (Figure 2.4(a)), des deux mélanges (Figure 2.4(b)) et des deux mélanges blanchis (Figure 2.4(c)).

effet la décorrélation n'est pas suffisante pour garantir la véracité de l'égalité (2.10), ce qui peut être vu grâce au développement de Taylor des fonctions caractéristiques des densités de probabilités conjointe et marginalesfaisant intervenir les cumulants croisés qui doivent être nuls pour assurer l'indépendance.

Ainsi, Comon [38] a proposé de généraliser l'analyse en composantes principales, qui n'impose l'indépendance qu'au second ordre et définit par conséquent des directions orthogonales, à l'*analyse en composantes indépendantes* (ACI) qui définit des directions indépendantes. Pour être performante, l'ACI devra donc recourir à des statistiques d'ordre supérieur (à deux). Ceci montre aussi pourquoi des sources gaussiennes iid ne peuvent être séparée. En effet, leur statistiques d'ordre supérieur à deux sont entièrement définies à partir de leur deux premières statistiques : l'utilisation des statistiques d'ordre supérieur n'apporte, dans ce cas particulier, aucune information supplémentaire.

Un autre moyen de montrer l'insuffisance de la décorrélation pour la séparation de sources est algébrique. Pour déterminer la matrice de séparation G de taille $(N_s \times N_s)$, en tenant compte des N_s indéterminations du gain, fixées de façon arbitraire, nous devons estimer $N_s^2 - N_s = N_s\,(N_s - 1)$ paramètres inconnus. Or les contraintes de décorrélation : $\mathrm{E}[z_i z_j] = 0$ pour toutes les paires $1 \leq i \neq j \leq N_s$, ne donnent que $N_s\,(N_s - 1)/2$ équations, ce qui est insuffisant pour déterminer G. Nous pouvons résumer ceci en disant que la décorrélation (indépendance à l'ordre deux) des sorties ne fait que "*la moitié de l'ACI*". Bien qu'insuffisante pour effectuer la séparation des sources, l'ACP permet, comme nous allons le voir, de simplifier le problème de l'ACI en contraignant la matrice de séparation G à adopter une structure particulière. Ainsi, pour achever la séparation par ACI, nous devons estimer une matrice U telle que

$$G = U\,W \tag{2.16}$$

soit une matrice séparante (*cf.* figure 2.5). L'indépendance des signaux estimés $\mathbf{y}$ implique aussi leur décorrélation : $C_{\mathbf{yy}} = \mathrm{E}\left[\mathbf{y}\mathbf{y}^T\right] = I_{N_s}$, en fixant de façon arbitraire la puissance moyenne des sources estimées à l'unité. Les signaux estimés étant définis

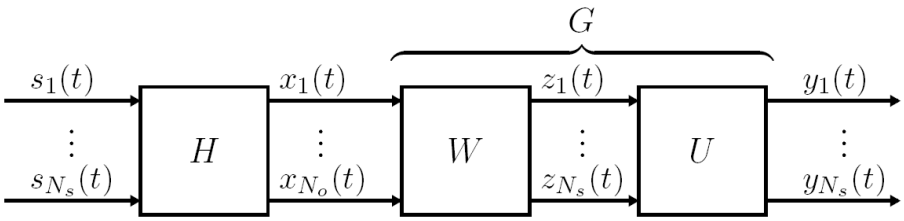


FIG. 2.5 – La décorrélation laisse une matrice de rotation inconnue.

par $\mathbf{y} = G\,\mathbf{x} = U\,\mathbf{z}$, nous avons

$$C_{\mathbf{yy}} = \mathrm{E}\left[\mathbf{y}\mathbf{y}^T\right] = U\,C_{\mathbf{zz}}\,U^T = U\,U^T$$

d'où nous déduisons que $U\,U^T = I_{N_s}$ traduisant le fait que la matrice U est une matrice unitaire ou matrice de rotation. Finalement, l'ACI peut se voir comme un pré-blanchiment (*i.e.* une ACP) suivi d'une rotation (figure 2.5).

Les méthodes de SAS peuvent se diviser en deux grandes catégories : d'une part les méthodes qui contraignent le processus de séparation en le décomposant en une étape de blanchiment spatial (ACP) suivi de la détermination de la matrice unitaire résiduelle et d'autre part les méthodes qui vont estimer directement le processus de séparation sans étape préliminaire de blanchiment. Pour la première catégorie de méthodes, les statistiques du second ordre sont utilisées pour la décorrélation tandis que les statistiques d'ordre supérieur vont permettre d'estimer la matrice résiduelle sous contrainte de son unitarité. Pour la seconde catégorie de méthodes, les statistiques d'ordre (deux et supérieur) sont exploitées directement pour l'estimation de la matrice de séparation sans imposer aucune contrainte à celle-ci (*i.e.* sans imposer la décorrélation des sources estimées).

Fonctions de contraste

La notion de *fonction de contraste* pour la séparation de sources a été introduite par Comon [38], qui s'est lui même inspiré des travaux de Donoho sur la déconvolution aveugle [48]. Une fonction de contraste $\Phi[\cdot]$, pour la séparation de sources, est une fonction à valeur réelle d'une densité de probabilité d'un vecteur aléatoire $\mathbf{s}$ qui est minimum[6] si ses composantes sont indépendantes. En d'autres termes, une fonction de contraste $\Phi[\cdot]$ vérifie

$$\Phi[C\,\mathbf{s}] \geq \Phi[\mathbf{s}] \tag{2.17}$$

pour tout vecteur aléatoire $\mathbf{s}$ à composantes indépendantes et toute matrice C. L'égalité n'est vérifiée que lorsque $C \cong I$. Ainsi, par exemple l'information mutuelle $I[\mathbf{y}]$ est une fonction de contraste [38] : $\Phi[\mathbf{y}] = I[\mathbf{y}]$. En effet, l'information mutuelle est minimale lorsque les composantes de $\mathbf{y}$ sont indépendantes traduisant de façon pratique la définition d'une fonction de contraste (2.17). Puisque la séparation de

[6]Selon la définition de Comon [38], le contraste est maximum à la séparation. Ici nous adoptons la convention opposée car de façon classique on cherche souvent à minimiser des critères.

sources peut être ramenée à la détermination d'une matrice de rotation en blanchissant les observations au préalable et en forçant les sorties $\mathbf{y}$ à être décorrélées, Cardoso [28] propose de définir les *fonctions de contraste orthogonal*, notées $\Phi^{\circ}[\cdot]$, comme des fonctions de contraste qui devront être minimisées sous la contrainte de décorrélation des sources estimées.

Ainsi, un algorithme de séparation de sources pourra être fondé sur la minimisation d'une fonction de contraste (orthogonal) appliquée aux sorties $\mathbf{y}$ de la fonction de séparation $\mathcal{G}(\cdot)$.

2.2.3 Séparation par mesure directe de l'indépendance

Nous allons maintenant présenter quelques-uns des principaux algorithmes de séparation de sources fondés sur une mesure directe de l'indépendance.

Méthode originelle

La première méthode proposée pour la séparation aveugle de source (l'algorithme de Hérault-Jutten [68]) est fondée sur l'utilisation de réseaux de neurones [66, 68, 76]. Leur principe pour rendre indépendantes deux variables aléatoires centrées y_1 et y_2 est d'utiliser deux fonctions impaires non linéaires $f(\cdot)$ et $g(\cdot)$ et de chercher à décorréler $f(y_i)$ et $g(y_j)$, pour $i \neq j$ (*i.e.* $\mathrm{E}[f(y_i)\, g(y_j)] = 0$). En effet, une condition nécessaire pour que la décorrélation soit réalisée est que tous les moments croisés impairs de y_1 et y_2 soient nuls ce qui fournit suffisamment d'équations de contraintes pour pouvoir estimer la matrice de séparation. L'algorithme de Hérault-Jutten est fondé sur une méthode adaptative d'annulation de $\mathrm{E}[f(y_i)\, g(y_j)] = 0$, où l'estimation des sources $\mathbf{y}(t)$ est définie par :

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{x}(t) - M\,\mathbf{y}(t) \\ &= (I + M)^{-1}\mathbf{x}(t). \end{aligned}$$

Par la suite d'autres méthodes fondées sur des réseaux de neurones ont été proposées [36]. Elles exploitent toutes la propriété des fonctions non-linéaires qui génèrent des termes d'ordre supérieur à deux. Ces moments croisés d'ordre supérieur à deux sont ensuite annulés par des algorithmes.

Maximum de vraisemblance

Le maximum de vraisemblance [92] est une technique classique du traitement du signal. Il permet d'estimer un ensemble θ de paramètres en maximisant la (log-)vraisemblance normalisée $L_T(\theta)$, définie comme (le logarithme de) la densité de probabilité, conditionnellement à θ, d'un ensemble T de réalisations $x(t)$, pour $1 \leq t \leq T$, d'une variable aléatoire :

$$L_T(\theta) \triangleq \frac{1}{T} \ln p_x[x(1), \cdots, x(T)|\theta].$$

Si l'on suppose les réalisations indépendantes (comme c'est le cas pour les processus aléatoires blancs), cette log-vraisemblance peut s'exprimer par

$$L_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \ln p_x[x(t)|\theta].$$

Appliqué à la séparation de sources [54, 100, 28], il consiste à estimer la matrice de séparation G à partir de la distribution des observations $\mathbf{x}(t) = H\,\mathbf{s}(t)$. Pour cela, la densité de probabilité $p_{\mathbf{x}}[\cdot]$ des observations peut être exprimée à partir de celle (supposée connue) des sources $p_{\mathbf{s}}[\cdot]$ par

$$p_{\mathbf{x}}[\mathbf{x}(t)|H] = |\det H^{-1}|\, p_{\mathbf{s}}[H^{-1}\,\mathbf{x}(t)], \tag{2.18}$$

en supposant que H est inversible. En se souvenant que G est une estimée de l'inverse de H, la log-vraisemblance normalisée $L_T(G)$ pour l'ensemble des T observations $\mathbf{x}(t)$, pour $1 \leq t \leq T$, noté $\{\mathbf{x}(t)\}_{1\leq t\leq T}$, s'écrit

$$L_T(G) = \frac{1}{T} \ln p_{\mathbf{x}}[\mathbf{x}(1), \cdots, \mathbf{x}(T)|G] = \frac{1}{T} \sum_{t=1}^{T} \ln p_{\mathbf{s}}[G\,\mathbf{x}(t)] + \ln|\det G|. \tag{2.19}$$

Le maximum de vraisemblance consiste donc à estimer la matrice de séparation G qui maximise $L_T(G)$. De plus, Cardoso [28] a montré que

$$L_T(G) \xrightarrow[T\to\infty]{} -KL[G\,\mathbf{x}\|\mathbf{s}] + cst$$

où, par abus de langage et dans un souci de ne pas alourdir les notations, $KL[\mathbf{x}\|\mathbf{y}]$ est la divergence de Kullback-Leibler entre les distributions de deux vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$. Ainsi, le maximum de vraisemblance est associé à la fonction de contraste

$$\Phi_{MV}[\mathbf{y}] \triangleq KL[\mathbf{y}\|\mathbf{s}]. \tag{2.20}$$

La log-vraisemblance normalisée $L_T(G)$ peut être vue alors comme une estimation de la fonction de contraste $\Phi_{MV}[\mathbf{y}]$ à une constante près. La minimisation de cette fonction de contraste (ou de façon équivalente la maximisation de la vraisemblance) nécessite la connaissance des densités de probabilité $p_{s_i}[\cdot]$ de chacune des sources s_i car les sources étant indépendantes, on peut écrire $p_{\mathbf{s}}[\cdot] = \prod_i p_{s_i}[\cdot]$. En pratique, ces densités de probabilités, qui sont inconnues dans un contexte aveugle, pourront être remplacées par celles estimées à partir des sources reconstruites $\mathbf{y}(t)$ en obtenant les mêmes performances asymptotiques que si les densités de probabilités des sources étaient connues à l'avance [101, 4]. Le maximum de vraisemblance peut ici s'interpréter comme suit [28] : "trouver une matrice de séparation G telle que la distribution de $\mathbf{y} = G\,\mathbf{x}$ soit aussi proche que possible, au sens de la divergence de Kullback-Leibler, de celle des sources $\mathbf{s}$".

Infomax

Une autre approche, introduite par Bell et Sejnowski [12], est, tout comme la méthode originelle, dérivée des réseaux de neurones. Elle est fondée sur la maximisation

de l'entropie $H[\mathbf{y}]$ des sorties $\mathbf{y}$, qui d'après (2.13) est égale à

$$H[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i] - I[\mathbf{y}].$$

Maximiser l'entropie jointe $H[\mathbf{y}] = H[G\,\mathbf{x}]$ par rapport à G revient à maximiser chacune des entropies marginales $H[y_i]$ et à minimiser $I[\mathbf{y}]$. Cependant cette maximisation n'est pas appropriée [26] et le principe infomax est alors de maximiser, par rapport à la matrice de séparation G, l'entropie $H[\mathbf{z}]$ de nouvelles variables aléatoires $\mathbf{z} = \mathbf{f}(\mathbf{y})$. Ici, $\mathbf{f}(\mathbf{y}) = [f_1(y_1), \cdots, f_{N_s}(y_{N_s})]^T$ est une fonction non-linéaire, composante à composante, telle que chaque z_i soit à distribution uniforme dans $[0, 1]$. Pour effectuer cette transformation, il suffit de choisir $f_i(\cdot) = \mathrm{P}_{y_i}[\cdot]$, où $\mathrm{P}_{y_i}[\cdot]$ est la fonction de répartition de y_i. Or d'après (2.13)

$$H[\mathbf{z}] = \sum_{i=1}^{N_s} H[z_i] - I[\mathbf{z}] = -I[\mathbf{z}] + cste,$$

la seconde égalité venant du fait que z_i est à distribution uniforme sur $[0, 1]$. Ainsi maximiser l'entropie conjointe $H[\mathbf{z}]$ revient à minimiser l'information mutuelle $I[\mathbf{z}]$ qui est minimum lorsque les composantes du vecteur aléatoire $\mathbf{z}$ sont indépendantes. La maximisation de $H[\mathbf{z}]$ permet bien d'obtenir l'indépendance des composantes de $\mathbf{y}$ puisque l'information mutuelle est invariante par transformation diagonale inversible (*i.e.* $I[\mathbf{z}] = I[\mathbf{f}(\mathbf{y})] = I[\mathbf{y}]$ si $\mathbf{f}(\cdot)$ est diagonale inversible, ce qui est le cas ici). Ainsi, le principe Infomax est associé à la fonction de contraste

$$\Phi_I[\mathbf{y}] \triangleq -H[\mathbf{f}(\mathbf{y})]. \tag{2.21}$$

Il est intéressant de noter que Cardoso [26] a montré que le principe Infomax était strictement équivalent au maximum de vraisemblance dans le cas de la séparation de sources $\Phi_I[\cdot] = \Phi_{MV}[\cdot]$. Appliquer ce principe aux réseaux de neurones revient à choisir les poids des neurones égaux aux coefficients de la matrice de séparation et les fonctions d'activation des neurones égales aux fonctions de répartition $\mathrm{P}_{y_i}[\cdot]$. Pour l'estimation de ces fonctions de répartitions, Bell et Sejnowski [12] propose un modèle paramétrique fondé sur la fonction logistique asymétrique généralisée qui permet de s'adapter à diverses allures de $\mathrm{P}_{y_i}[\cdot]$ (sur- et sous-gaussienne).

Minimisation de l'information mutuelle

L'approche la plus intuitive de la séparation de sources est certainement celle qui consiste à minimiser, par rapport à la matrice de séparation G, directement l'information mutuelle $I[\mathbf{y}]$ des sorties [74]. En effet, l'information mutuelle qui est une mesure directe de l'indépendance des composantes d'un vecteur, est une fonction de contraste [38]

$$\Phi_{IM}[\mathbf{y}] = I[\mathbf{y}]. \tag{2.22}$$

Notons que l'utilisation de l'information mutuelle peut être liée au maximum de vraisemblance [28]. En effet, soit $\tilde{\mathbf{y}}$ le vecteur aléatoire à composantes indépendantes

et tel que les densités de probabilité $\tilde{y}_i$ soient égales aux densités marginales de probabilité de y_i. Alors on a [39]

$$KL[\mathbf{y}\|\mathbf{s}] = KL[\mathbf{y}\|\tilde{\mathbf{y}}] + KL[\tilde{\mathbf{y}}\|\mathbf{s}].$$

Puisque $\mathbf{s}$ et $\tilde{\mathbf{y}}$ sont à composantes indépendantes, alors $KL[\tilde{\mathbf{y}}\|\mathbf{s}] = \sum_{i=1}^{N_s} KL[y_i\|s_i]$. Ainsi

$$\Phi_{MV}[\mathbf{y}] = \Phi_{IM}[\mathbf{y}] + \sum_{i=1}^{N_s} KL[y_i\|s_i] \tag{2.23}$$

où $KL[y_i\|s_i]$ mesure l'erreur de modélisation entre la densité de probabilité marginale des sorties y_i et celle choisie pour les sources. Il est alors possible d'interpréter le critère $\Phi_{MV}[\cdot]$ du maximum de vraisemblance en disant qu'il mesure l'écart total de modélisation comme étant la somme de l'écart des sources reconstitués à l'indépendance ($\Phi_{IM}[\cdot]$) plus l'écart de modélisation marginale ($\sum_{i=1}^{N_s} KL[y_i\|s_i]$). Remarquez que si les densités (marginales) de probabilité des sources sont estimées à partir des sources reconstituées y_i alors le deuxième terme de (2.23) s'annule et dans ce cas la méthode du maximum de vraisemblance et celle de la minimisation de l'information mutuelle sont équivalentes.

Nous venons de voir que maximum de vraisemblance, Infomax et information mutuelle sont intimement liés en séparation de sources. Nous allons voir qu'il est maintenant possible de les relier à la méthode originelle et de donner les fonctions non-linéaires optimales permettant de décrire au mieux l'indépendance [28].

En effet, si la matrice G permet la séparation des sources, elle doit annuler le gradient $\nabla\Phi_{IM}[G\,\mathbf{x}]$ de l'information mutuelle. Or $I[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i] - H[\mathbf{x}] - \ln|\det G|$, d'où

$$\nabla\Phi_{IM}[G\,\mathbf{x}] = \frac{\partial I[\mathbf{y}]}{\partial G} = \sum_{i=1}^{N_s} \frac{\partial H[y_i]}{\partial G} - \frac{\partial}{\partial G}\ln|\det G|. \tag{2.24}$$

Le gradient du second terme est simplement

$$\frac{\partial}{\partial G}\ln|\det G| = G^{-T} \tag{2.25}$$

tandis que

$$\begin{aligned}\frac{\partial H[y_i]}{\partial G_{ij}} &= -\mathrm{E}\left[\frac{\partial \ln p_{y_i}[y_i]}{\partial G_{ij}}\right] \\ &= \mathrm{E}\left[\psi_{y_i}[y_i]\,\frac{\partial y_i}{\partial G_{ij}}\right] \\ &= \mathrm{E}\left[\psi_{y_i}[y_i]\,x_j\right]\end{aligned}$$

où $\psi_{y_i}[y_i]$ est la fonction score marginale de la variable aléatoire y_i définie par

$$\psi_{y_i}[y_i] \triangleq -\frac{p'_{y_i}}{p_{y_i}}[y_i] \tag{2.26}$$

$(\cdot)'$ étant la dérivée. La fonction score représente toute l'information nécessaire sur la distribution de y_i. En regroupant ces résultats sous forme compacte on obtient

$$\nabla\Phi_{IM}[G\,\mathbf{x}] = \mathrm{E}\left[\Psi_{\mathbf{y}}\,\mathbf{x}^T\right] - G^{-T}$$

où $\Psi_{\mathbf{y}} = [\psi_{y_1}[y_1], \cdots, \psi_{y_{N_s}}[y_{N_s}]]^T$ est le vecteur des fonctions scores marginales. En multipliant à droite cette dernière expression par G^T et en exprimant le fait que ce gradient est nul à la séparation, on obtient les équations d'estimation

$$\mathrm{E}\left[\Psi_{\mathbf{y}}\,\mathbf{y}^T\right] - I = 0 \tag{2.27}$$

dont G est solution. Les équations $\mathrm{E}[\psi_{y_i}[y_i]\,y_i] = 1$, pour tout i, déterminent la puissance moyenne de la $i^{\text{ème}}$ source estimée. D'autre part, les équations $\mathrm{E}[\psi_{y_i}[y_i]\,y_j] = 0$, pour tout $i \neq j$, traduisent le fait que y_j doit être décorrélé d'une version non-linéaire $\psi_{y_i}[y_i]$ de la $i^{\text{ème}}$ source estimée. On retrouve l'idée originelle de Hérault et Jutten [68] de décorrélation non-linéaire mais ici le choix des fonctions non linéaires n'est plus heuristique. De plus, si les sources sont gaussiennes alors les fonctions scores sont linéaires et les équations d'estimation ci-dessus sont équivalentes à celles de décorrélation : les fonctions scores doivent être non linéaires pour faire apparaître les statistiques d'ordre supérieur nécessaires à la séparation.

La minimisation de ces fonctions de contraste (maximum de vraisemblance, Infomax ou information mutuelle) peut se faire par des techniques de type gradient en utilisant par exemple des algorithmes adaptatifs équivariants (EASI) [29]. La matrice de séparation est alors estimée de manière adaptative par

$$G_{t+1} = G_t - \alpha_t\,F(\mathbf{y}_t)\,G_t \tag{2.28}$$

où $F(\cdot)$ est obtenue par le gradient de la fonction de contraste que l'on cherche à minimiser et α_t est une suite de pas positifs. Ces méthodes ne nécessitent que la connaissance des fonctions scores marginales, comme nous l'avons vu ci-dessus. Celles-ci doivent être estimées puisqu'inconnues dans un contexte aveugle. Il existe diverses méthodes d'estimation : celles qui estiment les fonctions score de façon indirecte à partir des densités de probabilité en utilisant la définition (2.26) et celles qui vont estimer directement les fonctions score. Pour les méthodes indirectes, un principe classique d'estimation des densités de probabilité est d'utiliser un estimateur à noyaux [65] par exemple. Pour les méthodes directes, des approches paramétriques à base de fonctions non linéaires [101], de splines ou de polynômes [127, 97] par exemple ont été proposées.

2.2.4 Séparation par statistique d'ordre supérieur

Les méthodes de séparation par mesure directe de l'indépendance présentent l'avantage de modéliser de façon exacte l'indépendance des sources. Cependant, en pratique, ces techniques sont relativement lourdes à mettre en œuvre car elles requièrent à chaque mise à jour de la matrice de séparation (2.28) une nouvelle estimation des fonctions score marginales. Ainsi, d'autres techniques de séparation fondées sur les statistiques d'ordre supérieur, tels que les cumulants, ont été développées. Les cumulants sont définis comme étant les coefficients du développement de Taylor

de la seconde fonction caractéristique. Ils ont en outre l'avantage d'être additifs, multilinéaires et les cumulants d'ordre supérieur à deux sont nuls si la variable aléatoire est gaussienne. Bien que les méthodes de séparation fondées sur les cumulants puissent se justifier comme des approximations des critères de mesure directe de l'indépendance, on peut aussi les comprendre de façon heuristique. En effet, on a vu (*cf.* paragraphe 2.2.2) que la séparation de sources nécessite l'utilisation d'ordre supérieur pour avoir suffisamment d'équations pour estimer totalement la matrice de séparation G. L'emploi des critères que nous allons présenter permet d'ajouter, aux équations de décorrélation, de nouvelles contraintes nécessaires à la séparation des sources et nous verrons qu'ils définissent des fonctions de contraste justifiant ainsi leur emploi.

Maximisation de la non-gaussianité

En séparation de sources, le processus de mélange H a tendance à gaussianiser les sources : les observations $\mathbf{x} = H\,\mathbf{s}$ ont une distribution plus gaussienne que celle des sources $\mathbf{s}$. Ceci peut se justifier à partir du théorème limite central [92] : la distribution d'une somme de variables aléatoires indépendantes tend vers une distribution gaussienne. La séparation de sources ayant pour but de retrouver les sources (*i.e.* faire l'inverse du mélange), il semble naturel de chercher une matrice de séparation G telle que ces sorties $\mathbf{y}$ aient une distribution la moins gaussienne possible. Pour cela, la notion de non-gaussianité, qui permet de quantifier l'écart entre la distribution d'une variable aléatoire $\mathbf{x}$ et la variable aléatoire gaussienne $\mathbf{y}$ de même moyenne et même variance que $\mathbf{x}$, va être exploitée. Ceci peut se faire de deux façons : en utilisant d'une part le kurtosis et d'autre part la néguentropie.

Le kurtosis normalisé $\kappa[y]$ d'une variable aléatoire centrée y est défini comme le cumulant d'ordre quatre normalisé :

$$\kappa[y] \triangleq \frac{\mathrm{Cum}[y,y,y,y]}{\big(\mathrm{Cum}[y,y]\big)^2} = \frac{\mathrm{E}[y^4]}{(\mathrm{E}[y^2])^2} - 3. \tag{2.29}$$

Le kurtosis d'une variable aléatoire gaussienne est nul. Pour (presque) toutes les variables aléatoires non gaussiennes, le kurtosis est non nul. Puisque le kurtosis peut être positif ou négatif, une mesure de la non gaussianité est donnée par la valeur absolue de celui-ci $\sum_{i=1}^{N_s} |\kappa[y_i]|$ sous la contrainte de blanchiment des observations : $\mathrm{E}[\mathbf{y}\mathbf{y}^T] = I_{N_s}$. Mais dans ce cas, le critère n'est pas une fonction de contraste [28]. Pour obtenir une fonction de contraste à partir des kurtosis, il est nécessaire d'imposer l'hypothèse supplémentaire que toutes les sources ont un kurtosis de même signe. Dans ce cas, on peut définir une fonction de contraste orthogonal [70, 85] :

$$\Phi^{\circ}_{\kappa}[\mathbf{y}] = \sum_{i=1}^{N_s} \kappa[y_i] \tag{2.30}$$

si toutes les sources ont un kurtosis négatif. Si toutes les sources ont un kurtosis positif alors il suffit de prendre, pour fonction de contraste orthogonal, l'opposé de la formule ci-dessus. Cependant, l'estimation du kurtosis est, en pratique, sensible à la réalisation de la variable aléatoire [71] : le kurtosis n'est pas une mesure robuste de la non-gaussianité. Ainsi d'autres mesures de non-gaussianité ont été introduites.

La néguentropie $J[\cdot]$ est une autre mesure très courante de non-gaussianité. Elle est dérivée de la théorie de l'information [39] et on la définit par

$$J[\mathbf{y}] \triangleq H[\mathbf{y}_G] - H[\mathbf{y}] \tag{2.31}$$

où $\mathbf{y}_G$ est la variable aléatoire gaussienne de même moyenne et même variance que $\mathbf{y}$. La néguentropie est non négative et n'est nulle que si la variable aléatoire $\mathbf{y}$ est gaussienne. Cependant, la maximisation de ce critère par rapport à la matrice de séparation ne peut aboutir : en effet, la néguentropie étant invariante par transformation linéaire (*i.e.* $J[B\,\mathbf{y}] = J[\mathbf{y}]$ pour toute matrice inversible B) elle ne peut servir, sous sa forme conjointe, de critère de séparation puisque quelle que soit la matrice de séparation G on a $J[\mathbf{y} = G\,\mathbf{x}] = J[\mathbf{x}]$ indépendante de G ! Certains auteurs [71] proposent cependant d'utiliser la néguentropie mais sous sa forme marginale $J[y_i] = H[y_{g_i}] - H[y_i]$ et proposent alors de maximiser la néguentropie marginale de chaque sortie. La fonction de contraste associée est définie par :

$$\Phi_{neg}[\mathbf{y}] = -\sum_{i=1}^{N_s} J[y_i]. \tag{2.32}$$

Si l'on impose de plus la contrainte de décorrélation des sorties $\mathbf{y}$ de la matrice de séparation (*i.e.* la recherche d'une fonction de contraste orthogonal), et une fois blanchies les observations $\mathbf{z} = W\,\mathbf{x}$, comme expliqué au paragraphe 2.2.2, la séparation revient à chercher une matrice U unitaire à partir de $\mathbf{z}$. Il est alors possible de relier la minimisation de $\Phi_{neg}[\cdot]$ à la minimisation de l'information mutuelle sous contrainte d'orthogonalité $\Phi^{\circ}_{IM}[\cdot]$. Imposer $\mathrm{E}\left[\mathbf{y}\mathbf{y}^T\right] = I_{N_s}$, implique que $\Phi_{IM}[\mathbf{y}]$ (2.22) est, à une constante prés, égal à la somme des entropies marginales de chaque sortie y_i. En effet, $I[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i] - H[\mathbf{y}]$ par définition, or $\mathbf{y} = U\,\mathbf{z}$ d'où $I[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i] - H[\mathbf{z}] - \ln[\det U|$. Ainsi l'unitarité de U entraîne $\det U = 1$ et $I[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i] - H[\mathbf{z}]$. Le second terme du membre de droite est indépendant de U et il est possible de définir la fonction de contraste orthogonal suivante pour l'information mutuelle [28]

$$\Phi^{\circ}_{IM}[\mathbf{y}] = \sum_{i=1}^{N_s} H[y_i]. \tag{2.33}$$

Or $J[y_i] = H[y_{g_i}] - H[y_i]$, ainsi $\sum_i H[y_i] = \sum_i H[y_{g_i}] - \sum_i J[y_i]$ et d'après [69] on a $H[y_{g_i}] = \frac{1}{2}\ln\sigma_i^2 + \frac{1}{2}[1 + \ln(2\pi)]$ où σ_i^2 est la variance de y_{g_i} qui, sous contrainte de décorrélation des sources estimées $\mathbf{y}$ vaut 1. Donc $H[y_{g_i}]$ est indépendante de U. Finalement

$$\Phi^{\circ}_{IM}[\mathbf{y}] = cste + \Phi_{neg}[\mathbf{y}] \tag{2.34}$$

où *cste* est une constante indépendante de U, démontrant ainsi le lien étroit entre information mutuelle et néguentropie. Cependant l'utilisation directe de la néguentropie n'est pas aisée car elle nécessite le calcul des entropies faisant intervenir les densités de probabilité inconnues des sources reconstruites qu'il faut alors estimer. Ainsi, plusieurs approximations fondées sur divers développements des entropies ont été proposées [69].

En pratique les critères fondés sur le kurtosis (2.30) et sur la néguentropie (2.32), ou du moins une approximation de celle-ci, devront être minimisés. Pour cela, Hyvärinen et Oja ont proposé un algorithme rapide dit du point fixe (FastICA) [70, 69]

permettant une estimation des sources par une méthode de déflation orthogonale (*i.e.* estimation des sources les unes après les autres). Cette méthode est fondée sur le principe d'orthonormalisation de Gram-Schmidt. Une fois n sources estimées, la $(n+1)^{\text{ème}}$ suivante l'est par l'algorithme FastICA en s'assurant de plus qu'elle est orthogonale aux n précédentes.

Annulation des statistiques croisées d'ordre supérieur

Les méthodes fondées sur la non-gaussianité ont été introduites tout d'abord de façon heuristique à partir du théorème limite central et ont ensuite été reliées à l'information mutuelle qui demeure une mesure pratique de l'indépendance. Parallèlement, d'autres critères exploitant les statistiques d'ordre supérieur ont été dérivés directement de celle-ci.

Il est possible de montrer [38] que la fonction de contraste $\Phi_{IM}[\cdot]$ liée à l'information mutuelle peut être liée à la fonction de contraste orthogonal suivante

$$\Phi^{\circ}_{ICA}[\mathbf{y}] = \sum_{ijkl \neq iiii} \mathcal{C}^2_{ijkl}[\mathbf{y}] \tag{2.35}$$

où $\mathcal{C}_{ijkl}[\mathbf{y}] \triangleq \mathcal{C}[y_i, y_j, y_k, y_l]$ est un cumulant croisé d'ordre quatre. L'indépendance peut aussi être obtenue à partir d'un critère dérivé de $\Phi_{ICA}[\cdot]$, mais sur un sous-ensemble de cumulants plus restreints [28]

$$\Phi^{\circ}_{JADE}[\mathbf{y}] = \sum_{ijkl \neq ijkk} \mathcal{C}^2_{ijkl}[\mathbf{y}]. \tag{2.36}$$

Cette fonction de contraste a tout d'abord été obtenue par des considérations sur les matrices de cumulants diagonalisées conjointement [31]. Par la suite d'autres critères fondés sur les cumulants croisés d'ordre quatre ont été proposés [27]. Ces fonctions de contraste orthogonal, qui cherchent à minimiser des cumulants croisées d'ordre quatre, peuvent être également comprises d'un point de vue algébrique. En effet, comme nous l'avons rappelé, la contrainte de décorrélation n'est pas suffisante pour garantir l'indépendance qui nécessite l'annulation de tous les cumulants croisés d'ordre supérieur. Ces fonctions de contraste orthogonal, qui sont des approximations de cette condition, seront performantes pour estimer la matrice unitaire U si elles ajoutent suffisamment de contraintes (*i.e* annulent suffisamment de cumulants croisés d'ordre supérieur). En pratique la minimisation de ces fonctions de contraste orthogonal $\Phi^{\circ}_{ICA}[\cdot]$ et $\Phi^{\circ}_{JADE}[\cdot]$ peut être mise en œuvre par la technique de Jacobi [38] exploitant les matrices de rotations de Givens : la matrice unitaire U de taille $(N_s \times N_s)$ est déterminée par une séquence de matrices de rotation élémentaires de taille (2×2) appliquées à toutes les paires (y_i, y_j) pour $i \neq j$. L'angle de ces matrices de rotation est déterminé à chaque pas de façon analytique. Le critère $\Phi^{\circ}_{JADE}[\cdot]$ présente l'avantage de pouvoir déterminer à chaque pas les angles des matrices de rotation élémentaires de façon analytique même dans le cas de signaux complexes tandis que pour le critère $\Phi^{\circ}_{ICA}[\cdot]$, les angles ne sont déterminés de façon analytique que dans le cas de signaux réels.

2.2.5 Séparation semi-aveugle

Nous venons de présenter diverses stratégies pour effectuer la séparation de sources dans un cadre aveugle. Nous allons voir maintenant que si nous avons des connaissances *a priori* sur les sources, nous pouvons intégrer ces informations dans la détermination du processus de séparation H.

Sources non iid

Les méthodes présentées jusqu'ici ont été développées avec l'hypothèse (implicite ou explicite) que les sources sont iid. Nous avons ainsi vu que de telles méthodes ne peuvent pas se limiter à l'emploi de statistiques d'ordre deux (ce qui exclut donc les sources gaussiennes puisque totalement décrites par leurs statistiques d'ordre un et deux). Nous présentons maintenant deux nouvelles situations qui ont été proposées dans le cas où les sources sont supposées colorées (*i.e.* corrélées temporellement) c'est-à-dire si l'on lève le premier "i" de iid, puis dans le cas où les sources sont non stationnaires c'est-à-dire si l'on lève l'hypothèse "id" de iid.

Sources colorées Dans le cadre de sources colorées, plusieurs approches ont été proposées en étendant tout d'abord le principe du maximum de vraisemblance pour des sources avec un modèle autorégressif [44] ou grâce à un modèle d'état des sources avec un filtrage de Kalman [83]. Ces principes présentent l'inconvénient de recourir à des modèles paramétriques de sources qui devront être estimés ou fixés *a priori*. Ainsi, d'autres principes de séparation qui exploitent uniquement les statistiques d'ordre deux ont été proposés [14]. Après un blanchiment spatial W des observations, la séparation de sources se résume à une matrice unitaire U inconnue qui sera estimée de telle sorte que les matrices $C_{\mathbf{yy}}(\tau) \triangleq \mathrm{E}\left[\mathbf{y}(t)\mathbf{y}^T(t-\tau)\right]$ de fonctions d'autocorrélation des sources reconstruites $\mathbf{y}(t)$ soient diagonales pour tout τ. En effet les sources étant indépendantes et colorées, les matrices $C_{\mathbf{ss}}(\tau)$ de fonctions d'autocorrélation sont diagonales :

$$\forall \tau, \quad C_{\mathbf{ss}}(\tau) \triangleq \mathrm{E}\left[\mathbf{s}(t)\mathbf{s}^T(t-\tau)\right] = \mathrm{diag}(\gamma_1(\tau), \cdots, \gamma_{N_s}(\tau)) \tag{2.37}$$

où $\mathrm{diag}(\cdot)$ est la matrice diagonale dont les termes diagonaux sont les variables d'entrée et $\gamma_i(\tau) \triangleq \mathrm{E}[s_i(t)s_i(t-\tau)]$ est la fonction de covariance de la $i^{\text{ème}}$ source. La matrice U est donc déterminée de telle sorte que, pour un ensemble de décalages $\{\tau_k\}_{1\leq k\leq K}$ choisi, les matrices $C_{\mathbf{yy}}(\tau_k) = U\, C_{\mathbf{zz}}(\tau_k)\, U^T$ soient les plus diagonales possibles. L'estimation de la matrice U est faite par un algorithme de diagonalisation conjointe de l'ensemble des matrices $\{C_{\mathbf{yy}}(\tau_k) = U\, C_{\mathbf{zz}}(\tau_k)\, U^T\}_{1\leq k\leq K}$ grâce à la minimisation du critère

$$\mathcal{C}(U) = \sum_{k=1}^{K} \mathrm{off}\left(U\, C_{\mathbf{zz}}(\tau_k)\, U^T\right)$$

où $\mathrm{off}(A) = \sum_{1\leq i\neq j\leq N_s} |A_{ij}|^2$. La diagonalisation conjointe ne cherche pas à rendre diagonale chaque matrice $C_{\mathbf{zz}}(\tau_k)$ individuellement grâce à U, mais procède de telle sorte que l'ensemble des matrices $\{C_{\mathbf{zz}}(\tau_k)\}_k$ soit un ensemble de matrices presque diagonales. Plusieurs algorithmes d'identification aveugle au second ordre

(par exemple SOBI [14]) exploitent ce principe et permettent même d'extraire des sources gaussiennes pour peu qu'elles aient des spectres différents. En effet, ce principe permet la séparation des sources si l'ensemble $\{C_{\mathbf{yy}}(\tau_k)\}_{1\leq k\leq K}$ contient suffisamment de matrices $C_{\mathbf{yy}}(\tau_k)$ différentes, ce qui est le cas si les matrices $\{C_{\mathbf{ss}}(\tau_k)\}_{1\leq k\leq K}$ sont différentes.

Sources non stationnaires De la même manière, relâcher la nature id des sources conduit à considérer des sources non stationnaires et à exploiter des algorithmes de séparation exploitant uniquement les statistiques d'ordre deux [15, 35, 99]. Les sources étant indépendantes et non stationnaires, les matrices de covariance $C_{\mathbf{ss}}(t)$ des sources sont diagonales (dû à l'indépendance) et différentes en fonction de t (dû à la non-stationnarité). Ainsi, en divisant l'intervalle $\mathcal{T} = \{t\}_t$ de tous les instants d'observation t en K sous-intervalles consécutifs (ou recouvrant partiellement) $\mathcal{T}_k$ ($\mathcal{T} = \bigcup_k \mathcal{T}_k$), on définit la matrice de covariance $C_{\mathbf{xx}}(\mathcal{T}_k) \triangleq \mathrm{E}_{t\in\mathcal{T}_k}[\mathbf{x}(t)\mathbf{x}(t)^T]$ pour $t \in \mathcal{T}_k$. L'estimation de la matrice de séparation H est faite par un algorithme de diagonalisation conjointe de l'ensemble des matrices $\{C_{\mathbf{yy}}(\mathcal{T}_k) = H\, C_{\mathbf{xx}}(\mathcal{T}_k)\, H^T\}_{1\leq k\leq K}$ grâce à la minimisation du critère

$$\mathcal{C}(H) = \sum_{k=1}^{K} w_k \,\mathrm{off}\left(H\, C_{\mathbf{xx}}(\mathcal{T}_k)\, H^T\right)$$

avec $w_k = |\mathcal{T}_k|/|\mathcal{T}|$, où $|\mathcal{T}|$ est le cardinal de l'ensemble $\mathcal{T}$. Ce critère qui se comprend de façon intuitive est aussi lié au maximum de vraisemblance et à l'information mutuelle [99]. Le choix de la partition $\{\mathcal{T}_k\}_k$ va influencer l'estimation des matrices de covariance $C_{\mathbf{xx}}(\mathcal{T}_k)$ et donc la qualité de la séparation. Dans [95] par exemple, la variance à court terme des sources est supposée constante par morceaux sur une partition $\{\mathcal{T}_k\}_k$ fixée *a priori* et telle que tous les $\mathcal{T}_k$ soient de même longueur. Une autre possibilité [98, 30] est d'optimiser et d'ajuster la taille des $\mathcal{T}_k$ en effectuant un compromis entre finesse de la partition de $\mathcal{T}$ (*i.e.* taille de chaque $\mathcal{T}_k$) et complexité de la partition (*i.e.* nombre de sous-intervalles K).

Une autre approche, fondée elle aussi sur l'utilisation des statistiques d'ordre deux, exploite la diversité temps-fréquence des sources [15, 13] modélisée par les matrices de la distribution spatiale temps-fréquence des signaux définies par

$$D_{\mathbf{ss}}(t,f) = \sum_{l=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \phi(m,l)\,\mathbf{s}(t+m+l)\mathbf{s}^T(t+m-l)\,e^{-j4\pi fl}$$

où le noyau $\phi(m,l)$ caractérise la distribution. La matrice unitaire U est estimée de telle sorte qu'elle diagonalise conjointement l'ensemble $\{D_{\mathbf{zz}}(t_k,f_k)\}_{1\leq k\leq K}$ des matrices $D_{\mathbf{zz}}(t_k,f_k)$ des observations blanchies $\mathbf{z}(t)$.

Ainsi, il est intéressant de remarquer que, bien que les statistiques d'ordre deux ne soient pas suffisantes pour la séparation des sources iid (*cf.* paragraphe 2.2.2), dès lors que les sources possèdent une structure temporelle différente (coloration ou non-stationnarité) il est possible de les séparer en exploitant uniquement les statistiques d'ordre deux. En conséquence, il devient alors possible de séparer les sources gaussiennes. En effet, il suffit d'introduire suffisamment d'équations de contrainte (au

moins $N_s(N_s-1)/2$) pour que l'on puisse déterminer tous les coefficients de la matrice de séparation H. La séparation de sources colorées et/ou non stationnaires fait intervenir des critères de séparation fondés sur des diagonalisations conjointes de matrices qui peuvent être effectuées, par exemple, en utilisant l'algorithme JADE [31] ou celui proposé par Pham [96].

Sources parcimonieuses

La parcimonie des sources est une autre information *a priori* très performante pour la séparation de sources. Elle consiste à dire que dans une certaine base de représentation, par exemple le plan temps-fréquence ou dans le domaine de la transformée en cosinus discrète (TCD), les sources ont une représentation parcimonieuse : les sources peuvent être non stationnaires et s'éteindre par moment (c'est par exemple le cas s'un signal de parole spontané) ou n'occuper qu'une partie du spectre ou encore combiner les deux.

Une première approche [3, 2, 1] exploite la parcimonie des signaux dans le plan temps-fréquence. Dans le cas de deux sources et deux observations (pour plus de généralité se référer à [1]), le rapport entre les deux observations dans le plan temps-fréquence (t,f) donne

$$\alpha(t,f)=\frac{X_1(t,f)}{X_2(t,f)}=\frac{H_{11}S_1(t,f)+H_{12}S_2(t,f)}{H_{21}S_1(t,f)+H_{22}S_2(t,f)} \tag{2.38}$$

où $X(t,f)$ est la transformée de Fourier à court terme (TFCT) de $x(t)$. Pour estimer la matrice de mélange H (ou du moins la matrice de mélange à un facteur d'échelle près), il suffit de trouver des zones temps-fréquence (t_k,f_l) où une seule source est active. En effet, si la première (resp. deuxième) source est inactive dans la zone temps-fréquence (t_k,f_l), alors $\alpha(t_k,f_l)=H_{12}/H_{22}$ (resp. $\alpha(t_k,f_l)=H_{11}/H_{21}$) ce qui permet d'estimer la matrice de mélange, colonne par colonne, au facteur d'échelle près $\mathrm{diag}(H_{11},H_{12})$. Cette méthode repose donc sur la capacité à détecter des zones temps-fréquence où une seule source est active, ce qui peut se faire en étudiant la variance du rapport $\alpha(t,f)$ qui est minimale si au plus une source est active. En effet si au plus une source est active alors le rapport $\alpha(t,f)$ devient déterministe (défini par les coefficients de matrice de mélange) est donc de variance théorique nulle tandis que si au moins deux sources sont actives, ce rapport est aléatoire donc de variance non nulle. Cependant, bien que l'étude de la variance du rapport $\alpha(t,f)$ permette de déterminer les zones temps-fréquence où au plus une source est active, celle-ci ne permet pas d'identifier de quelle source il s'agit. Pour cela, les auteurs proposent d'utiliser un critère sur la distance entre les valeurs du rapport $\alpha(t,f)$ qui sont différentes suivant la source active. Notez que l'utilisation du plan temps-fréquence n'est pas strictement nécessaire pour cette méthode : le même principe peut être envisagé dans le domaine temporel ou tout autre espace de représentation obtenu par transformation linéaire. Cependant, l'activation d'au plus une source dans le domaine temporel est une hypothèse plus restrictive que celle de l'activation d'au plus une source dans une zone temps-fréquence particulière ce qui réduit d'autant le champ des signaux éligibles. Cette méthode peut être vue comme une généralisation de l'algorithme DUET [72, 139] qui repose sur l'hypothèse de W-disjointe-orthogonalité

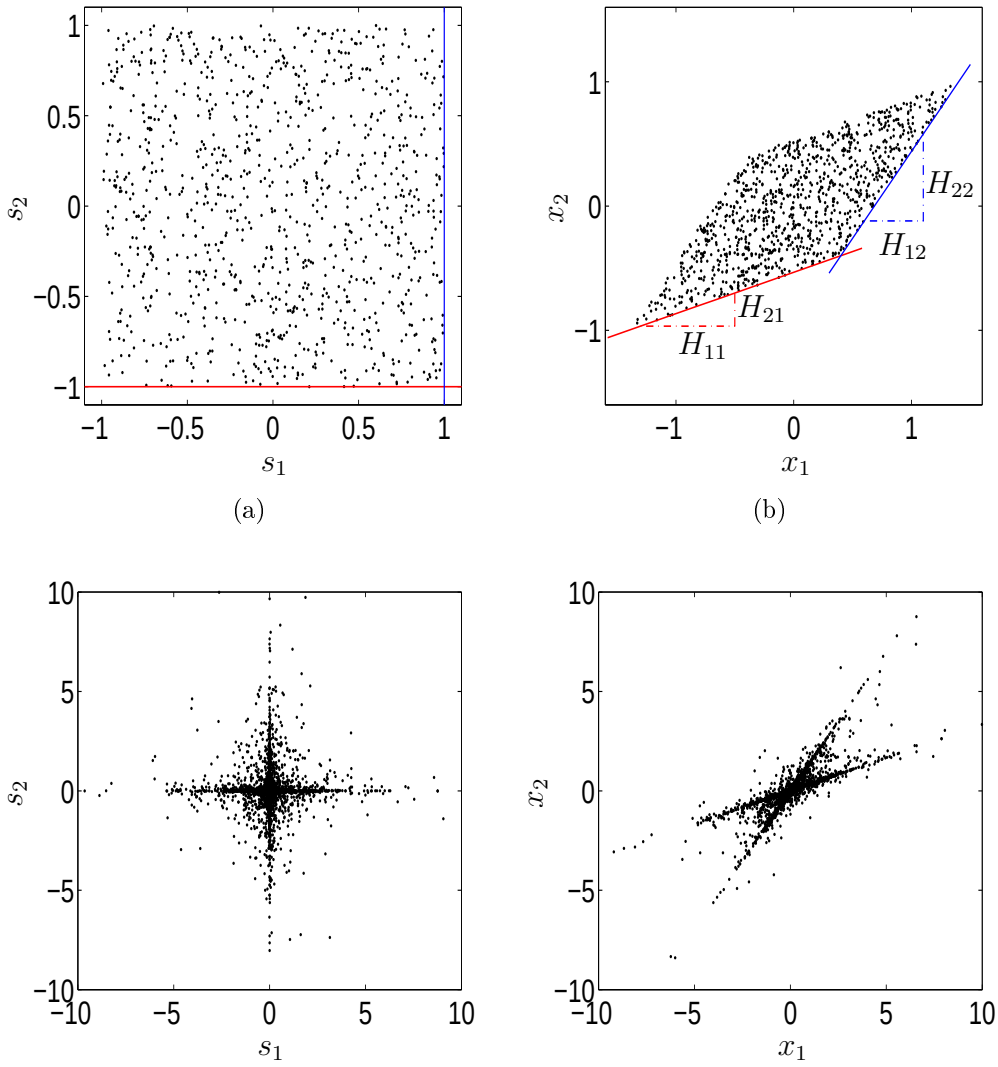
dans le plan temps-fréquence associé à la transformée considérée avec une fenêtre de pondération W. Cette hypothèse revient à supposer que les sources sont à support disjoint dans le plan temps-fréquence : elles ne présentent alors aucun recouvrement et il est possible d'extraire les sources par un simple masquage. Cette hypothèse apparaît plus forte et donc plus restrictive que celle utilisée par Abrard et Deville [1]. D'une façon plus générale, d'autres études [62, 75, 63] exploitent la parcimonie des sources en effectuant une transformation linéaire telle que dans la nouvelle représentation seul un nombre limité de coefficients ait une amplitude significative. Le problème de la séparation de source est alors d'identifier la matrice de mélange puis d'estimer la représentation des sources avant de les reconstruire.

La seconde approche de séparation de sources exploitant la parcimonie des sources que nous présentons est fondée sur des considérations géométriques [8, 47]. La première méthode de séparation géométrique [106] exploite le fait que pour des sources à distribution ayant des bords nets (comme c'est le cas pour la distribution uniforme *cf.* figure 2.6(b)) l'équation des bords de la distribution jointe des mélanges donne la direction des sources dans les mélanges. En estimant ces directions, il est possible d'estimer la matrice de mélange H à un facteur d'échelle près : les coefficients directeurs des bords de la distribution jointe des mélanges $\mathbf{x}(t)$ sont donnés par H_{21}/H_{11} et H_{22}/H_{12}. Dans le cas de sources parcimonieuses, celles-ci ont des distributions piquées (*i.e.* concentrées autour de leur valeur moyenne) et les bords ne sont pas nets (*cf.* figure 2.6(c)) : cette méthode n'est plus directement applicable. Cependant, les directions des sources sont toujours nettement visibles dans les mélanges comme le montre la figure 2.6(d). Il donc encore possible de les estimer. Pour cela, [8] propose une méthode fondée sur une analyse en composantes principales par classes. Les observations $\mathbf{x}(t)$ sont divisées en autant de classes que de sources et on calcule la direction principale de chacune des classes. Celles-ci sont alors mises à jour en redistribuant les données en fonction de la distance des observations $\mathbf{x}(t)$ aux droites définies par les directions principales déterminées précédemment et ainsi de suite jusqu'à convergence. Les coefficients des équations des directions principales ainsi déterminées correspondent aux colonnes de la matrice de mélange H à un facteur près. Pour achever la séparation, il suffit alors d'inverser la matrice H. Notons que cette méthode repose sur l'hypothèse que toutes les sources sont parcimonieuses.

Il est intéressant de remarquer que les méthodes de séparation de sources parcimonieuses peuvent être étendues au cas de mélanges sous-déterminés. En effet, que ce soit par la parcimonie dans le plan temps-fréquence ou dans une autre représentation ou encore par des considérations géométriques, ces méthodes permettent l'estimation de la matrice de mélange H. Cependant, même si la matrice de mélange a pu être identifiée, elle ne permet pas l'estimation des sources car cette matrice n'est pas inversible : dans le cas sous-déterminé, l'estimation de la matrice de mélange est un problème différent de celui de l'estimation des sources.

2.3 Mélanges convolutifs

Considérons maintenant le cas des mélanges linéaires convolutifs dans lequel, le processus de mélange $\mathcal{H}(\cdot)$ est supposé *linéaire* et *avec mémoire* (*cf* Figure 2.7). Les N_s sources (statistiquement indépendantes) sont donc supposées être mélangées



(a) (b)

(c) Sources (d) Mélanges

FIG. 2.6 – Séparation géométrique. Distributions conjointes : de deux sources indépendantes uniformes (resp. parcimonieuses) figure 2.6(a) (resp. figure 2.6(c)) et des deux mélanges correspondants (Figures 2.6(b) et 2.6(d)).

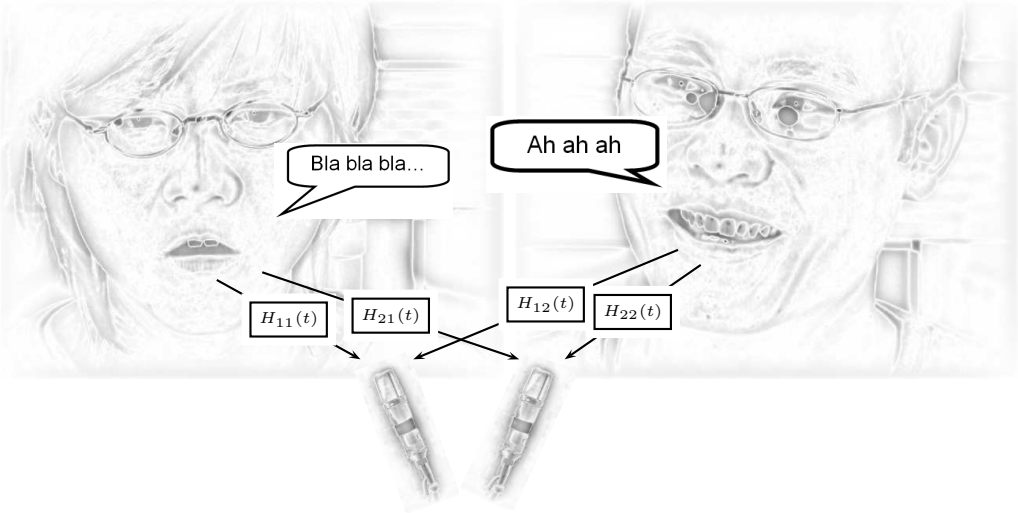


FIG. 2.7 – *Cocktail party* dans le cadre de mélanges linéaires convolutifs : les canaux de transmission entre les sources et les capteurs sont modélisés par des filtres de réponse impulsionnelle $H_{i,j}(t)$.

convolutivement et linéairement : chacune des N_o observations peut ainsi s'exprimer par

$$x_i(t) = \sum_{j=1}^{N_s} h_{i,j}(t) * s_j(t), \quad \forall i \in \{1, \cdots, N_o\}. \tag{2.39}$$

Ceci traduit le fait que chaque observation $x_i(t)$ est une combinaison linéaire des sources $s_j(t)$ filtrées par des filtres de réponse impulsionnelle $h_{i,j}(t)$. On peut réécrire ce système sous forme matricielle

$$\mathbf{x}(t) = H(t) * \mathbf{s}(t) \tag{2.40}$$

en faisant apparaître $H(t)$ la *matrice de filtres de mélange*, de dimension $N_o \times N_s$, qui a pour $(i,j)^{\text{ème}}$ élément la réponse impulsionnelle $h_{i,j}(t)$. La séparation de sources consiste alors à estimer une *matrice de filtres de séparation* $G(t)$, de dimension $N_s \times N_o$, telle que ses sorties

$$\mathbf{y}(t) = G(t) * \mathbf{x}(t) \cong \mathbf{s}(t) \tag{2.41}$$

soient des estimées des sources originales $\mathbf{s}(t)$. En d'autres termes, la matrice $\mathbf{G}(t)$, dont les réponses impulsionnelles pourront estimées grâce à l'hypothèse d'indépendance des sources, doit être une matrice séparante.

2.3.1 Séparabilité et indéterminations

Il a été prouvé [138] que les mélanges convolutifs sont séparables : c'est-à-dire que l'indépendance des composantes de $\mathbf{y}(t)$ (2.41) assure la séparation des sources. Ainsi, chercher des signaux mutuellement indépendants est équivalent à retrouver les sources. Tout comme pour les mélanges instantanés, l'indétermination de permutation demeure : l'indépendance des sources estimées ne dépend pas de l'ordre dans

lequel on les reconstruit. De plus, si pour les mélanges instantanés l'indétermination du facteur d'échelle n'a d'autre conséquence que de retrouver les sources à un gain constant près, dans le cas de mélanges convolutifs, cette indétermination diagonale $\Lambda(\cdot)$ (*cf.* paragraphe 2.1.2) devient maintenant un filtre d'indétermination : les sources ne pourront être estimées qu'à une permutation globale près et à un filtre près. Cette indétermination de facteur d'échelle peut être inacceptable puisqu'elle peut entraîner une très forte distorsion des signaux estimés (les rendant inaudibles ou fortement dégradé dans le cas de signaux auditifs par exemple).

Pour résoudre le problème de séparation de sources de mélanges convolutifs, deux grandes catégories de méthodes ont été proposées : d'une part celles qui cherchent à estimer les filtres de séparation dans le domaine temporel (*i.e.* estimation des réponses impulsionnelles des filtres, *cf* paragraphe 2.3.2) et d'autre part celles qui les estiment dans le domaine fréquentiel (*i.e.* estimation des réponses en fréquence des filtres, *cf* paragraphe 2.3.3).

2.3.2 Séparation temporelle

La première approche proposée pour les mélanges convolutifs le fut par Jutten *et al.* [73]. Leur algorithme se propose de séparer deux sources à partir de deux observations. Les filtres à réponse impulsionnelle finie sont estimés à partir d'une généralisation de leur critère proposé pour les mélanges instantanés (*cf.* paragraphe 2.2.3). Ils cherchent à annuler les corrélations non linéaires

$$E\left[f(s_i(t))\, g(s_j(t-\tau))\right] = 0 \tag{2.42}$$

où $f(\cdot)$ et $g(\cdot)$ sont deux fonctions impaires fixées *a priori* [73] ou optimisées en même temps que la recherche des filtres de séparation [32]. Mais l'annulation de ces corrélations non linéaires s'avère délicate. Il est alors possible d'exploiter des idées semblables à celles développées dans le cadre de mélanges instantanés : l'utilisation de statistiques d'ordre supérieur ou des méthodes de séparation semi-aveugle reposant sur des hypothèses *a priori* faites sur les sources. Ainsi, il a été prouvé [90] que l'indépendance de $\mathbf{y}(t)$ peut être obtenue à partir de critères fondés sur les cumulants d'ordre quatre, comme par exemple $\mathrm{Cum}[y_i(t), y_i(t), y_j(t-\tau), y_j(t-\tau)]$ ou $\mathrm{Cum}[y_i(t), y_j(t-\tau), y_j(t-\tau), y_j(t-\tau)]$ ou encore $\mathrm{Cum}[y_i(t), y_i(t), y_i(t), y_j(t-\tau)]$. De même, la non-stationnarité peut être utile pour permettre la séparation de façon performante des sources [136, 84]. Ainsi l'annulation des corrélations croisées $E[y_1(t)y_2(t-\tau)]$ et $E[y_1(t-\tau)y_2(t)]$ pour différents retards τ permet de générer suffisamment d'équations de contraintes pour pouvoir effectuer la séparation de sources en ne requérant que des statistiques d'ordre deux.

Une autre approche pour l'estimation des filtres de séparation est fondée sur l'annulation des bi-spectres croisés des sorties $\mathbf{y}(t)$ du processus de séparation $\mathcal{G}(\cdot)$ [138]. Les polyspectres, qui sont en fait une représentation fréquentielle des cumulants, d'un ensemble de K processus $x_1(t), \cdots, x_K(t)$ associé aux indices $k_0, \cdots, k_m \in \{1, \cdots, K\}$ sont définis par

$$P_{x_{k_0}, x_{k_1} \cdots, x_{k_m}}(\omega_1, \cdots, \omega_m) \triangleq \mathrm{TF}\left(\mathrm{Cum}[x_{k_0}(t), \cdots, x_{k_m}(t+\tau_m)]\right) \tag{2.43}$$

où $\mathrm{TF}(\cdot)$ est l'opérateur transformée de Fourier. Ainsi, l'annulation conjointe des bi-spectres $P_{y_1^*, y_1, y_2}(\omega_1, \omega_2)$ et $P_{y_2^*, y_2, y_1}(\omega_1, \omega_2)$ implique la séparation des sources.

Bien que relativement performantes, ces méthodes n'en demeurent pas moins complexes à mettre en œuvre et sont coûteuses en temps de calcul notamment en raison du nombre de paramètres important requis par les filtres de séparation. Ainsi, d'autres méthodes de séparation fondées sur le domaine fréquentiel ont été proposées.

2.3.3 Séparation fréquentielle

L'estimation des filtres de séparation $G(t)$ directement dans le domaine temporel est, comme nous venons de le voir, un problème délicat : il s'agit de résoudre un problème convolutif. En exploitant la propriété de la transformée de Fourier (TF) qui dit que la transformée de Fourier d'un produit de convolution est égale au produit des transformées de Fourier (*i.e.* $TF(a*b) = TF(a)\,TF(b)$), il est possible de reformuler le problème de séparation différemment. En supposant de plus que les processus de mélanges $H(t)$ et $G(t)$ sont stationnaires (*i.e.* n'évoluant pas au cours du temps) les équations de mélange (2.40) et de séparation (2.41) s'expriment

$$\mathbf{X}(t,f) = H(f)\,\mathbf{S}(t,f) \quad (2.44)$$

$$\mathbf{Y}(t,f) = G(f)\,\mathbf{X}(t,f) \quad (2.45)$$

où $\mathbf{Z}(t,f)$ est la transformée de Fourier à court terme (TFCT) de $\mathbf{z}(t)$ et $H(f)$ (resp. $G(f)$) est la transformée de Fourier des filtres de mélanges $H(t)$ (resp. séparation $G(t)$). Le fait que les fonctions de mélange $H(t)$ et de séparation $G(t)$ soient supposés stationnaires implique que leurs réponses en fréquence $H(f)$ et $G(f)$ ne dépendent pas du temps. Ainsi, à chaque fréquence f_i, ces nouvelles équations s'identifient à un problème instantané. Le passage dans le domaine fréquentiel permet donc de transformer la résolution d'un problème de séparation de sources de mélange convolutif en N_f problèmes de séparation de sources de mélange instantané, où N_f est le nombre de fréquences de calcul de la TFCT. Pour estimer les réponses en fréquence $G(f)$ des matrices de séparation, il suffit d'appliquer une méthode adaptée aux mélanges instantanés (*cf.* paragraphe 2.2) et ce à chaque fréquence f_i de calcul de la TFCT. Cependant, la transformée de Fourier a tendance à gaussianiser les signaux, et notamment les sources $\mathbf{S}(t,f)$, ce qui implique qu'il est préférable de recourir à des méthodes de séparation de sources semi-aveugle (*cf.* paragraphe 2.2.5) qui n'exploitent que les statistiques d'ordre deux : par exemple la coloration ou la non-stationnarité des sources [93, 40, 102] ou la parcimonie dans le plan temps-fréquence [108, 139, 105].

Les méthodes fondées sur l'hypothèse de non-stationnarité des sources recherchent une matrice de séparation $G(f)$ qui diagonalise conjointement un ensemble de matrices $\{\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k,f)\}_{\mathcal{T}_k}$ de densités spectrales de puissance à court terme des observations $\mathbf{X}(t,f)$, où $\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k,f) \triangleq E[\mathbf{X}(t,f)\mathbf{X}^+(t,f)]$ avec $t \in \mathcal{T}_k$ un ensemble prédéfini d'indices temporels et $^+$ est le conjugué transposé. En effet, les sources étant indépendantes, $\Gamma_{\mathbf{SS}}(t,f) \triangleq E[\mathbf{S}(t,f)\mathbf{S}^+(t,f)]$ est une matrice diagonale. Ainsi, pour achever la séparation, une idée naturelle est de chercher, à chaque fréquence f_i, une matrice de séparation $G(f_i)$ telle que les matrices $\{G(f_i)\,\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k,f_i)\,G^+(f_i)\}_{\mathcal{T}_k}$ soient diagonales. Para et Spence [93] utilisent cette idée avec le critère

$$\mathcal{C}(G(f)) = \sum_{k=1}^{K} \left\| G(f)\,\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k,f)\,G^+(f) - \mathrm{diag}(G(f)\,\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k,f)\,G^+(f)) \right\|^2 \quad (2.46)$$

qui est minimisé par un algorithme du type gradient. En dérivant le principe de l'information mutuelle (ou de façon équivalente le maximum de vraisemblance), Pham *et al.* [102] propose de minimiser le critère suivant

$$\mathcal{C}(G(f)) = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \det \operatorname{diag}(G(f)\, \Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k, f)\, G^{+}(f)) - \ln \det G(f) \right\} \qquad (2.47)$$

en exploitant un algorithme rapide de diagonalisation conjointe [96]. Cependant ces techniques souffrent d'un inconvénient majeur : l'indétermination de permutation. En effet, il n'est possible d'estimer la matrice de séparation $G(f)$ qu'à une permutation $\Pi(f)$ près (en omettant la distorsion diagonale) :

$$\forall f,\ \exists \Pi(f)\ /\ G(f) = \Pi(f)\, H^{-1}(f). \qquad (2.48)$$

La séparation étant effectuée à chaque fréquence indépendamment des autres fréquences, rien n'assure que les matrices $\{\Pi(f)\}_f$ soient les mêmes pour toutes les fréquences f comme illustré à la figure 2.8. Il est tout à fait envisageable d'obtenir des permutations différentes à chaque fréquence : par blocs de fréquences consécutives ou de façon isolée (*cf.* figures 2.8(a) et 2.8(b)). Sans régularisation des permutations, on obtient des sources indépendantes sans pour autant avoir satisfait la séparation des sources : il est nécessaire d'introduire un post-traitement après la séparation de façon à résoudre les indéterminations de permutations rencontrées à chaque fréquence. Ainsi le post-traitement doit permettre d'assurer que

$$\forall (f_1, f_2), \quad \Pi(f_1) = \Pi(f_2). \qquad (2.49)$$

Notons que, même si l'on peut régulariser les permutations, il restera éventuellement une permutation globale (*i.e.* la même permutation inconnue à chaque fréquence). Pour cela plusieurs approches ont été proposées en faisant de nouvelles hypothèses sur les fonctions de mélanges [93, 102] ou sur les sources [115]. Ainsi, imposer que les réponses en fréquence $G_{ij}(f)$ des filtres de séparation soient à variation douce en fonction de f implique que les réponses impulsionnelles $G_{ij}(t)$ soient courtes : [93] impose donc comme contrainte lors de la minimisation de (2.46) que les coefficients $G(t) = 0$ pour $t > Q$, où Q est la longueur de la réponse impulsionnelle des filtres de séparation, ce qui lient les fréquences entre elles. Exploitant la même idée de régularité locale des réponses en fréquence des filtres de séparation, Pham *et al.* [102] proposent d'initialiser la diagonalisation conjointe de $\{\Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k, f_i)\}_{\mathcal{T}_k}$ par la matrice de séparation estimée à la fréquence f_{i-1} : à la fréquence f_i, ils cherchent donc la matrice $D(f_i)$ telle qu'elle diagonalise conjointement $\{G(f_{i-1})\, \Gamma_{\mathbf{X},\mathbf{X}}(\mathcal{T}_k, f_i)\, G^{+}(f_{i-1})\}_{\mathcal{T}_k}$ et ainsi $G(f_i) = D(f_i)G(f_{i-1})$. Bien que ces méthodes apportent une régularisation correcte des permutations, celles-ci sont mises en défaut dès lors que l'hypothèse de variations lentes des réponses en fréquence des filtres n'est plus vérifiée, ce qui est le cas si les filtres de mélange présentent de nombreux échos. Ainsi, [115] propose un critère fondé sur l'évolution temporelle des éléments diagonaux des matrices $\Gamma_{\mathbf{Y},\mathbf{Y}}(t, f)$ de densité spectrale à court terme des signaux estimés $\mathbf{y}(t)$. Le $k^{\text{ème}}$ élément diagonal de $\Gamma_{\mathbf{Y},\mathbf{Y}}(t, f)$ représente la variation de l'énergie spectrale de la $k^{\text{ème}}$ source estimée à la fréquence f au cours du temps t. Les auteurs appellent *profil*

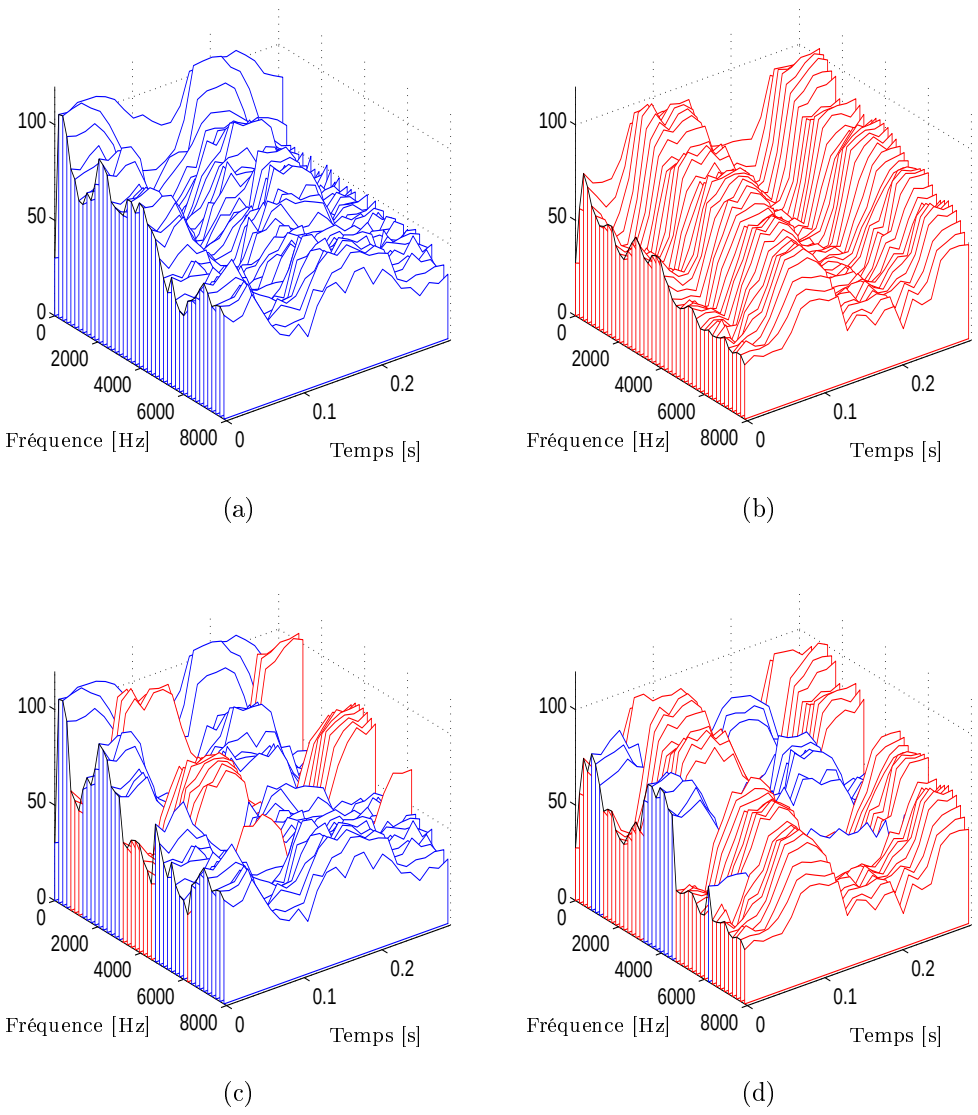


FIG. 2.8 – Problème des permutations pour la séparation fréquentielle : densité spectrale de puissance à court terme $\Gamma_{ss}(t, f)$ de deux sources indépendantes ([UPUP] 2.8(a) et [APAPA] 2.8(b)) et des deux sources estimées avec des permutations nuisibles (2.8(c) et 2.8(d)).

le logarithme de cette valeur et le notent $E(f,t;k)$. L'idée est que, pour une même source et pour des fréquences voisines, l'évolution des profils est similaire comme le montre la figure 2.8. Cependant l'indétermination du facteur d'échelle ne permet pas l'utilisation directe des profils : ils ne sont estimés qu'à une constante additive près. En effet, les profils sont le logarithme de l'énergie spectrale des sources estimées qui le sont à une constante multiplicative près (due à l'indétermination du facteur d'échelle). Ainsi, les auteurs définissent les *profils centrés* comme les profils auxquels on retranche leur moyenne temporelle :

$$E'(f,t;k) = E(f,t;k) - \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} E(f,t;k),$$

où $\mathcal{T}$ est l'ensemble des indices temporels de calcul des TFCT. Les profils centrés sont donc indépendants du facteur d'échelle inconnu. Le principe de régularisation des permutations est alors de chercher l'ensemble des permutations $\{\Pi(f)\}_f$ tel que les profils centrés $E'(\cdot,t;k)$ (ici considérés comme une fonction de la fréquence) soient à variations douces par rapport à la fréquence f.

Finalement, si le fait de transformer la séparation d'un mélange convolutif en la séparation de plusieurs mélanges instantanés semble simplifier la séparation en permettant l'emploi de techniques utilisées pour les mélanges instantanés, cela déplace en fait le problème : il est nécessaire de régulariser les permutations ce qui n'est pas aisé.

Parallèlement, d'autres approches exploitant la parcimonie dans le plan temps-fréquence ont été proposées [108, 139, 105] dans le cadre de mélanges convolutifs simplifiés se résumant à une atténuation et un retard temporel : les filtres de mélanges sont alors de la forme $H_{ij}(t) = H_{ij}\,\delta(t-\tau_{ij})$ où H_{ij} et τ_{ij} sont l'atténuation et le retard associés au canal entre la $j^{\text{ème}}$ source et le $i^{\text{ème}}$ capteur et $\delta(t)$ l'impulsion de Dirac. Ces méthodes sont en fait des généralisations des algorithmes de séparation semi-aveugle de source de mélange instantané exploitant la parcimonie dans le plan temps-fréquence (*cf.* paragraphe 2.2.5). Il est ici nécessaire d'estimer l'atténuation et le retard de chacun des filtres ce qui peut être fait en étudiant le rapport $X_i(t,f)/X_1(t,f)$ [108, 139, 105], puisqu'en fixant de façon arbitraire les filtres $H_{1,j}(t) = 1$ pour tous j, on obtient

$$\frac{X_i(t,f)}{X_1(t,f)} = H_{ij}\,e^{-\jmath 2\pi\tau_{ij}f}$$

si seulement la $j^{\text{ème}}$ source $S_j(t,f)$ est présente. Comme dans le cas instantané, [108, 139] supposent que les sources sont W-disjointes-orthogonales tandis que [105] fait une recherche des zones temps-fréquence où une seule source est active.

2.4 Séparation de sources de parole audiovisuelle

La séparation de sources trouve de nombreuses applications notamment en réhaussement de la parole avec le problème de la *cocktail party* : il s'agit alors de séparer un ou plusieurs signaux de parole parmi d'autres signaux concurrents de parole et de bruit ambiant. Si l'utilisation des signaux audio est chose courante,

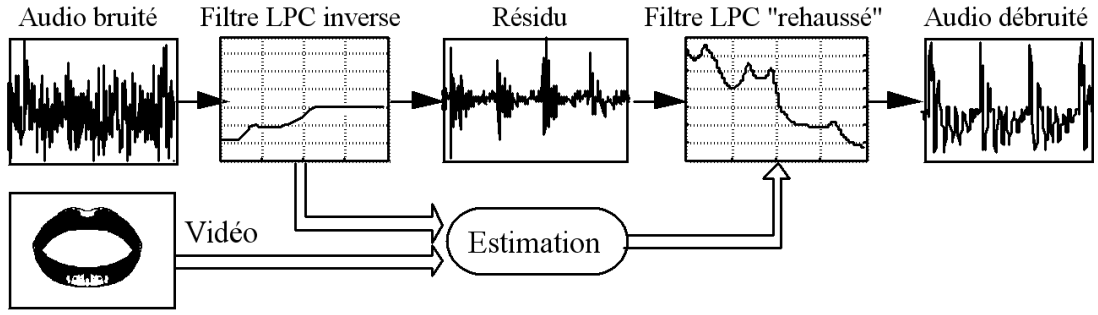


FIG. 2.9 – Débruitage de Wiener audiovisuel.

l'adjonction de la modalité visuelle (*i.e.* le visage des locuteurs) est encore peu utilisée. De plus, contrairement à la séparation de sources, qui cherche à retrouver toutes les sources, dans le problème de la *cocktail party*, on focalise sur un ou plusieurs locuteurs spécifiques dont le signal de parole est considéré comme signal d'intérêt. Ainsi, le problème de séparation de sources de parole peut s'apparenter à celui du réhaussement de parole et nous parlerons plus facilement d'extraction de sources de parole.

Le premier principe de réhaussement de parole audiovisuelle fut proposé par Girin *et al.* [59]. Son principe est d'effectuer un filtrage de Wiener sur le signal bruité $x(t) = s(t) + n(t)$ où $s(t)$ est le signal de parole non bruité, $n(t)$ est un bruit blanc gaussien et $x(t)$ est l'observation bruité. Notons que ce problème s'apparente à celui de la séparation de sources de mélange instantané sous-déterminé de deux sources $s(t)$ et $n(t)$ à partir d'une seule observation $x(t)$. Le filtrage de Wiener consiste à appliquer un filtre de réhaussement (ou de séparation) dont la réponse en fréquence $H(f)$ est telle que

$$H(f) = \frac{\Gamma_s(f)}{\Gamma_x(f)} \tag{2.50}$$

où $\Gamma_s(f)$ (resp. $\Gamma_x(f)$) est la densité spectrale de puissance de $s(t)$ (resp. $x(t)$). L'estimation de $\Gamma_x(f)$ ne posant pas de problème puisque le signal bruité $x(t)$ est connu, la modalité visuelle intervient dans l'estimation de $\Gamma_s(f)$ qui n'est pas directement accessible. En effet, si l'on dispose des paramètres des lèvres du locuteur, il est possible de construire une relation entre ceux-ci et les paramètres du modèle de prédiction linéaire (LP) du signal audio utilisé alors pour fournir une estimation de $\Gamma_s(f)$. Un filtre de séparation dérivé de (2.50) peut alors se mettre sous la forme [59]

$$H(z) = \frac{1 + \sum_{1=1}^{p} b_{x,i}\, z^{-1}}{1 + \sum_{1=1}^{p} \hat{b}_{s,i}\, z^{-1}}$$

où p est l'ordre du modèle LP, $\{b_{x,i}\}_i$ sont les coefficients du modèle LP correspondant à l'observation bruitée $x(t)$ et $\{\hat{b}_{s,i}\}_i$ sont les coefficients estimés du modèle LP du signal de parole. Ces derniers sont estimés par régression linéaire entre les paramètres vidéo et les paramètres audio du signal bruité. Le principe du débruitage audiovisuel proposé est résumé à la figure 2.9.

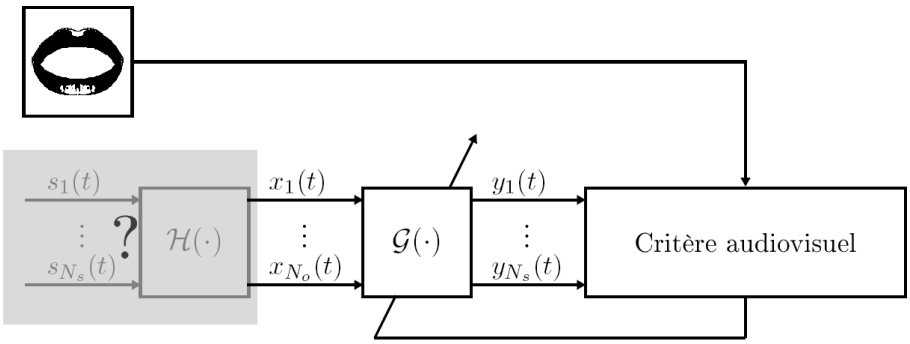


FIG. 2.10 – Séparation de sources audiovisuelle.

Développant l'idée originelle de la bimodalité de la parole, Sodoyer *et al.* [120, 118, 117] proposent un système d'extraction de source de parole fondé non plus sur un modèle linéaire liant les paramètres audio et vidéo mais sur un modèle statistique. L'idée de l'extraction est de reconstruire, à l'aide des observations, le signal le plus cohérent possible avec les paramètres vidéo disponibles (*cf.* figure 2.10). Dans le cadre de mélanges instantanés, les observations $\mathbf{x}(t)$ sont liées aux sources $\mathbf{s}(t)$ par une matrice de mélange H (*cf.* paragraphe 2.2). Pour extraire la source $s_1(t)$ liée au signal vidéo $v_1(t)$, il suffit d'estimer la première ligne[7] $G_{1:}$ de la matrice de séparation G : $y_1(t) = G_{1:}\mathbf{x}(t)$. Pour cela, [120] propose de maximiser la cohérence entre le signal estimé $y_1(t)$ et le signal vidéo $v_1(t)$ grâce au modèle audiovisuel statistique $p_{AV}(y_1(t), v_1(t))$. Ainsi, on peut définir comme critère audiovisuel à minimiser

$$\mathcal{C}_{AV}(G_{1:}) = -\ln p_{AV}(y_1(t), v_1(t)). \tag{2.51}$$

Récemment, en exploitant de la même manière la cohérence audiovisuelle, [134] étend ce principe au cas des mélanges convolutifs et propose aussi d'exploiter la cohérence audiovisuelle comme fonction de contrainte pour la séparation [135].

2.5 Conclusion

Dans ce chapitre, nous venons d'introduire la séparation de sources qui consiste à estimer des sources inconnues à partir de mélanges de celles-ci. La formalisation de ce problème en terme d'analyse en composantes indépendantes, au début de années 1990, a conduit la communauté du traitement du signal à proposer de nombreux principes de séparation dans un cadre aveugle exploitant uniquement l'hypothèse d'indépendance mutuelle des sources. Ces algorithmes exploitent soit une mesure directe et exhaustive de l'indépendance au prix d'un fort coût de calcul, soit une mesure indirecte et partielle de l'indépendance amenant à une complexité algorithmique moindre. Depuis, la séparation semi-aveugle de sources a été envisagée : celle-ci tient compte de spécificités *a priori* (connues ou supposées) des sources

[7] Les ":" indiquent que la dimension correspondante dans une matrice est vectorisée. Ainsi $A_{i:}$ est le vecteur ligne correspondant à la $i^{\text{ème}}$ ligne de la matrice A : $A_{i:} = [A_{i1}, \cdots, A_{in}]$ et $A_{:j}$ est le vecteur colonne regroupant les éléments de la $j^{\text{ème}}$ colonne de A : $A_{:j} = [A_{1j}, \cdots, A_{nj}]^T$.

pour proposer de nouveaux algorithmes de séparation. Pour sa part, la séparation de sources de parole audiovisuelle est un domaine naissant et encore largement sous exploré. Les travaux pionniers dans ce domaine donnent des résultats prometteurs. Mais ceux-ci ne demeurent encore qu'au stade académique des mélanges instantanés et ne fonctionnent que pour des corpus simplistes.

Deuxième partie

Modélisation de la multimodalité de la parole

Chapitre 3

Modèle audiovisuel de la parole

Beaucoup d'applications en traitement du signal (par exemple la compression [57] ou le réhaussement de parole [51]) requièrent une connaissance statistique du signal. Il en est de même dans le problème de la séparation de source. On a vu par exemple, au chapitre 2, une série de méthodes exploitant la densité de probabilité *a priori* des sources, qu'il est alors nécessaire d'estimer ou de modéliser. Dans notre approche audiovisuelle de la séparation de source de parole, nous développerons au chapitre 5 une première technique qui exploite la cohérence et la complémentarité des signaux audiovisuels de la parole sous la forme d'un modèle probabiliste. En effet, dans la communauté du traitement de la parole audiovisuelle, il est classique de considérer que la relation entre les paramètres vidéo et audio de la parole est complexe et peut être exprimée par une relation statistique [137, 120]. Cette complexité est en partie due au fait qu'à une forme de lèvres donnée, il n'est pas possible d'associer de façon univoque un son et donc une forme spectrale particulière. Par exemple, il est impossible à partir de la vision des lèvres de faire la distinction entre un [u] et un [y] (*cf.* paragraphe 1.2). Nous présentons dans ce chapitre les bases d'une telle modélisation statistique.

Dans un premier temps, nous chercherons à définir un modèle statistique *purement audio* bien adapté à notre approche fréquentielle de la séparation de sources qui, comme on le verra dans ce chapitre, nécessite de modéliser le comportement de coefficients de la transformée de Fourier discrète (TFD) à chaque fréquence de calcul. Généralement, la distribution de ces coefficients est compliquée. C'est pourquoi l'utilisation de modèles généraux est une solution largement répandue. En particulier, les mélanges de (plusieurs) noyaux gaussiens sont des modèles très utilisés pour les applications de traitement de la parole (par exemple, conversion de voix [124], réhaussement de la parole [24] ou identification du locuteur [107]). De tels modèles généraux offrent une bonne adéquation des données au modèle, mais au prix d'un nombre élevé de paramètres (multiplication des noyaux) et d'un fort coût calculatoire.

Par ailleurs, la parole est un signal non stationnaire : elle contient du silence et une large gamme de sons différents. En conséquence, la distribution d'un grand nombre de trames consécutives de parole continue peut être considérée comme pi-

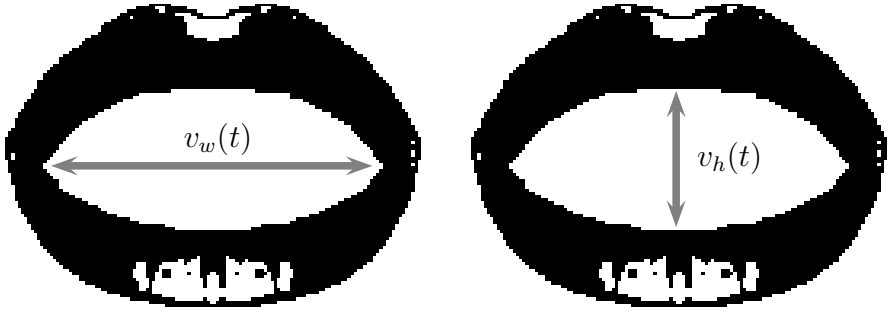


FIG. 3.1 – Paramètres vidéo : largeur interne et hauteur interne des lèvres.

quée[1]. Ainsi, un modèle *global* de la distribution du signal de parole, par exemple laplacien [86, 56], peut être utilisé. Mais ce modèle global ne tient pas compte du fait qu'une longue séquence de parole est composée de plusieurs sons. Ainsi, dans notre étude, nous allons exploiter la structure de la parole pour construire un modèle à base de noyaux tels que chacun d'entre eux modélise un son. De plus, nous proposons un tel modèle à noyaux nécessitant moins de paramètres et plus efficace qu'un modèle général multi-gaussien.

Dans un deuxième temps, nous étendons notre modélisation purement audio vers un modèle audiovisuel exploitant ces résultats. C'est ce modèle que nous exploiterons dans notre approche audiovisuelle de la séparation de sources de parole au chapitre 5.

Pour répondre à cette série d'objectifs, ce chapitre est organisé comme suit. Nous présentons tout d'abord les paramètres audiovisuels que nous allons utiliser tout au long de notre étude avant d'introduire un modèle multi-gaussien général. Nous verrons ensuite que nous pouvons concevoir un modèle spécifiquement adapté aux paramètres audio utilisés, nous donnant ainsi un modèle audiovisuel plus efficace que le modèle multi-gaussien. Enfin, nous présenterons les corpus utilisés avant de donner des résultats expérimentaux de notre modélisation.

3.1 Paramètres audiovisuels

Dans ce paragraphe, nous décrivons tout d'abord les paramètres visuels et les paramètres acoustiques que nous utilisons pour modéliser la bimodalité de la parole. Ce sont ces paramètres que nous exploiterons ensuite dans notre problématique de séparation de sources.

3.1.1 Paramètres visuels

Nous savons (paragraphe 1.2) que les lèvres véhiculent la majeure partie de l'information visuelle utile pour la compréhension de la parole. De plus, une paramétrisation exploitant uniquement la hauteur et la largeur intérolabiales est suffisante pour traduire la variance de l'information visuelle : le vecteur $\mathbf{v}(t)$ des paramètres

[1]Les zones de silence conduisent à avoir une densité de probabilité avec une densité forte au voisinage de 0.

vidéo regroupe donc $v_w(t)$ et $v_h(t)$, respectivement la largeur interne et la hauteur interne du contour labial (Figure 3.1)

$$\mathbf{v}(t) = \begin{pmatrix} v_w(t) \\ v_h(t) \end{pmatrix}. \tag{3.1}$$

Pour extraire ces paramètres labiaux, une micro-caméra, fixée sur la tête des locuteurs, est orientée vers la zone labiale (Figure 3.2). Les lèvres du locuteur sont maquillées en bleu de façon à ce qu'un système semi-automatique de traitement (*chroma-key*) puisse les segmenter facilement [80]. Cette segmentation permet ensuite de déterminer les paramètres vidéo utilisés dans notre étude en exploitant des algorithmes de suivi de contour. Les caméras utilisées sont à entrelacement et enregistrent 25 images par seconde. En désentrelaçant les lignes paires et impaires, nous obtenons de nouvelles images à la fréquence de 50 trames par seconde. Les paramètres vidéo peuvent ainsi être vus comme un signal échantillonné à 50Hz, fournissant des paramètres vidéo $\mathbf{v}(t)$ toutes les 20ms.

3.1.2 Paramètres audio

En traitement de la parole, il est courant d'utiliser le domaine spectral plutôt que le domaine temporel directement. Dans le cadre de notre étude, ceci est renforcé par le fait que la bimodalité de la parole lie l'information visuelle du visage du locuteur au spectre d'amplitude des sons émis[2]. Le vecteur $\mathbf{a}(t)$ des paramètres audio va regrouper des caractéristiques spectrales locales du signal acoustique de parole $s(t)$ correspondant au signal vidéo enregistré. Ainsi, $s(t)$ est divisé en trames consécutives de façon synchrone avec le signal vidéo. L'intervalle de temps entre deux trames consécutives est donc de 20ms puisque les modèles que nous allons présenter ont pour but d'associer chaque vecteur de paramètres audio à un vecteur de paramètres vidéo $\mathbf{v}(t)$. Cependant, que se passe-t-il si le locuteur prononce deux fois le même son mais avec des intensités différentes ? Dans ce cas, nous obtenons deux fois le même spectre mais avec des puissances différentes. Pour réduire la complexité des corpus, nous allons considérer que deux sons identiques mais à des puissances différentes sont les mêmes. Pour cela, chaque trame audio est centrée, normalisée et multipliée par une fenêtre de Hamming de façon à calculer la transformée de Fourier à court terme (TFCT) $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T$:

$$\mathbf{S}(t) = \mathrm{TF}\left(\frac{s_{N_f}(t) - \mathrm{E}[s_{N_f}(t)]}{\mathrm{Var}[s_{N_f}(t)]^{1/2}} w(t) \right) \tag{3.2}$$

où $s_{N_f}(t)$ est le signal $s(t)$ multiplié par une fenêtre rectangulaire de taille N_f localisée à l'instant t et $w(t)$ est la fenêtre de Hamming de longueur N_f. C'est donc l'étape de normalisation qui permet d'assurer que deux sons identiques à des puissances différentes produisent les mêmes coefficients $\mathbf{S}(t)$. Dans notre étude, les signaux acoustiques sont échantillonnés à 16kHz et nous choisissons des fenêtres d'analyse sans

[2]On entend par là que la forme des lèvres est liée à la répartition de l'énergie du signal en fonction de la fréquence mais elle ne peut pas être liée à une information précise de phase : le déphasage précis d'un son tenu ne peut pas être connu à partir de la forme des lèvres.

(a)

(b)

(c)

(d)

(e)

(f)

FIG. 3.2 – Enregistrement des paramètres vidéo. Figures 3.2(a), 3.2(b) 3.2(c) et 3.2(d) : conditions d'enregistrement. Figures 3.2(e) et 3.2(f) images de la zone des lèvres enregistrées par les micro-caméras.

recouvrement et de durée 20ms : $N_f = 320$. Les coefficients de la TFCT sont largement utilisés parce qu'ils produisent une représentation relativement parcimonieuse et efficace des signaux. Ainsi, ces coefficients spectraux, ou plutôt le logarithme de leur module, vont servir de base pour calculer le vecteur des paramètres audio $\mathbf{a}(t)$ utilisé dans les modèles audiovisuels. L'emploi du spectre en échelle logarithmique est justifié d'une part par des considérations de perception de l'oreille humaine [141] et d'autre part par un bon conditionnement des valeurs numériques des modules des coefficients spectraux.

Nous allons maintenant présenter deux modèles audiovisuels permettant de relier les paramètres vidéo $\mathbf{v}(t)$ et les paramètres audio $\mathbf{a}(t)$ calculés à partir du logarithme du module des coefficients de la TFCT.

3.2 D'un modèle audiovisuel général...

Dans un premier temps, nous avons repris le modèle multi-gaussien proposé par Sodoyer *et al.* [120]. Dans cette étude, le nombre de coefficients spectraux $\mathbf{a}(t)$ est réduit et différent du nombre de fréquences sur lesquelles on calcule la TFCT. Pour cela, des bancs de filtres permettent d'intégrer (*i.e.* moyenner) les coefficients $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T$ de la TFCT sur des bandes de fréquences consécutives à un instant t. Cette opération est représentée par la matrice $B \in \mathbb{R}^{N_b \times N_f}$, où N_f est le nombre de fréquences de calcul de la TFCT et N_b le nombre de bandes de fréquences. Ce moyennage fréquentiel est suivi d'une analyse en composantes principales (ACP) réalisée sur le logarithme du résultat. Les paramètres audio $\mathbf{a}(t)$ sont donc donnés par

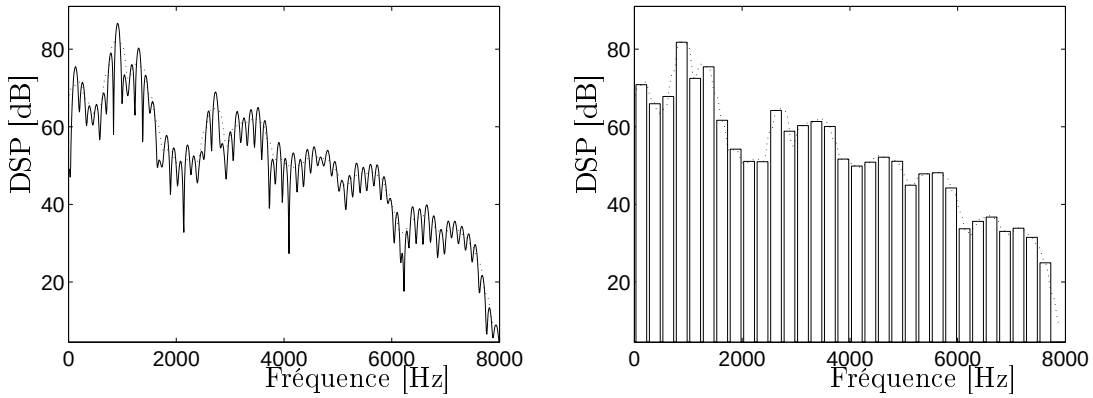
$$\mathbf{a}(t) = C \, \ln\left(B \, |\mathbf{S}(t)|^2\right), \tag{3.3}$$

où $C \in \mathbb{R}^{N_b \times N_b}$ est la matrice représentant l'ACP et $\ln(\mathbf{x})$ est le logarithme d'un vecteur défini composante à composante : $\ln(\mathbf{x}) = [\ln(x_1), \cdots, \ln(x_N)]^T$. Dans cette étude, les filtres composant le banc de filtres B sont des filtres passe-bande idéaux sans recouvrement. L'étape de moyennage fréquentiel permet de limiter la dimension du vecteur des paramètres audio tout en conservant l'allure spectrale des sons, *i.e.* les caractéristiques spectrales telles que les formants (Figure 3.3). Parallèlement, cette opération permet de s'affranchir des variations spectrales d'un son donné dues aux variations de la fréquence fondamentale. En général, lors de l'ACP, on fait une réduction de dimension supplémentaire en ne gardant que certaines composantes, celles qui représentent le plus de variance.

Sodoyer *et al.* [120] proposent, pour modéliser la relation entre les paramètres audio et vidéo, d'utiliser un modèle multi-gaussien (MMG). Dans ce cas, la probabilité conjointe audiovisuelle $p_{AV}(\cdot)$ est définie par

$$p_{AV}(\mathbf{a}(t), \mathbf{v}(t)) = \sum_{i=1}^{N_{AV}} \omega_i^{AV} \, p_G(\mathbf{z}_{AV}(t) | \mu_i^{AV}, \Sigma_i^{AV}) \tag{3.4}$$

où $\mathbf{z}_{AV}(t) = [\mathbf{a}^T(t), \mathbf{v}^T(t)]^T$ est le vecteur des paramètres audiovisuels, $\{\omega_i^{AV}\}_i$ est l'ensemble des poids des N_{AV} noyaux audiovisuels. $p_G(\cdot|\mu, \Sigma)$ est la densité de pro-



(a) (b)

FIG. 3.3 – Influence du banc de filtres. Figure 3.3(a) : densité spectrale de puissance d'un [a] (le trait pointillé correspond à la moyenne mobile sur 250Hz). Figure 3.3(b) : densité spectrale de puissance après le banc de filtres passe-bande idéaux sans recouvrement de largeur 250Hz (le trait pointillé correspond à la moyenne glissante sur 250Hz).

babilité normale sachant le vecteur des valeurs moyennes μ et la matrice de covariance Σ. Cette densité est définie d'une façon générale par

$$p_G(\mathbf{z}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right)$$

où d est la dimension du vecteur $\mathbf{z}$ et $\det(\cdot)$ est le déterminant. Ainsi, pour chaque noyau i du modèle (3.4), le vecteur audiovisuel des valeurs moyennes μ_i^{AV} est composé de la concaténation du vecteur des valeurs moyennes audio μ_i^A et de celui des valeurs moyennes vidéo μ_i^V :

$$\mu_i^{AV} = \begin{bmatrix} \mu_i^A \\ \mu_i^V \end{bmatrix}.$$

La matrice de covariance audiovisuelle Σ_i^{AV} se décompose quant à elle sous la forme

$$\Sigma_i^{AV} = \begin{pmatrix} \Sigma_i^A & \vdots & \left(\underline{\Sigma}_i^{AV}\right)^T \\ \cdots & \cdots & \cdots \\ \underline{\Sigma}_i^{AV} & \vdots & \Sigma_i^V \end{pmatrix}$$

où Σ_i^A et Σ_i^V sont respectivement les matrices de covariance des paramètres audio et vidéo du $i^{\text{ème}}$ noyau, $\underline{\Sigma}_i^{AV}$ est la matrice représentant les intercorrélations entre les paramètres audio et vidéo. L'ensemble $\{\omega_i^{AV}, \mu_i^{AV}, \Sigma_i^{AV}\}_i$ des paramètres du modèle audiovisuel (3.4) doit être estimé, ce qui peut être effectué par l'algorithme EM (*cf.* annexe B). Si, comme on escompte dans ce modèle, chaque noyau gaussien représente un son de parole, alors le vecteur des valeurs moyennes audiovisuelles μ_i^{AV} permet d'associer entre elles formes moyennes labiales et spectrales (*i.e.* énergie spectrale associée à une bande de fréquences) d'un seul son de parole. La matrice de covariance Σ_i^{AV} modélise alors la variabilité des paramètres audiovisuels d'un son de parole et la corrélation entre ces paramètres.

3.3 . . . vers un modèle audiovisuel spécifique

Le modèle (3.4) à noyaux gaussiens liant les paramètres vidéo (3.1) aux paramètres spectraux réduits (3.3) s'est montré efficace dans le cadre de la séparation de sources de mélanges instantanés [120, 118]. Cependant, il ne peut pas être utilisé tel quel (*i.e.* sur des énergies spectrales audio associées à des bandes de fréquences) dans le cadre de la séparation de sources fréquentielle de mélanges convolutifs que nous allons mettre en œuvre (*cf.* partie III) et ceci pour deux raisons. En effet, d'après le paragraphe 2.3.3, la séparation de sources fréquentielle de mélange convolutif consiste à estimer les filtres de séparation $H_{ij}(f)$, ce qui nécessite une connaissance *a priori* pour *toutes les fréquences de calcul* f de la TFCT. Cela induit d'une part de ne faire ni moyennage fréquentiel (*i.e.* $B = I_{N_f}$), ni ACP (*i.e.* $C = I_{N_f}$). Les coefficients audio $\mathbf{a}(t)$ sont alors définis par $\mathbf{a}(t) = \ln\left(|\mathbf{S}(t)|^2\right)$ ou à un facteur $1/2$ près par

$$\mathbf{a}(t) = \ln|\mathbf{S}(t)|. \tag{3.5}$$

Et d'autre part, pour chaque son, la distribution $p_i^A(\cdot)$ de ces nouveaux paramètres acoustiques $\mathbf{a}(t)$ est obtenue par marginalisation de chacun des noyaux audiovisuels gaussiens de (3.4) vis-à-vis des paramètres vidéo, $\forall i$:

$$\begin{aligned} p_i^A(\mathbf{a}(t)) &= \int p_G\big(\mathbf{z}_{AV}(t)\,\big|\mu_i^{AV}, \Sigma_i^{AV}\big)\, d\mathbf{v} \\ &= p_G\big(\mathbf{a}(t)\,\big|\mu_i^A, \Sigma_i^A\big)\,. \end{aligned}$$

Ce noyau purement acoustique est lui-même gaussien, de vecteur des valeurs moyennes μ_i^A et de matrice de covariance Σ_i^A. Or pour des sections de parole quasi-stationnaires (*i.e.* correspondant alors à un seul son), les coefficients complexes $\mathbf{S}(t)$ de la TFD peuvent être considérés comme ayant une distribution complexe gaussienne circulaire [103]. Ainsi, la distribution des paramètres audio (3.5), définis comme le logarithme du module de ces coefficients, $\ln|\mathbf{S}(t)|$, ne peut plus être gaussienne. Choisir des noyaux gaussiens pour modéliser la relation statistique entre les paramètres audio (3.5) et vidéo (3.1) ne sera donc pas adaptée aux paramètres à modéliser. Par conséquent, une telle modélisation sera peu efficace puisque plusieurs noyaux gaussiens seront alors nécessaires pour modéliser chaque son.

Par conséquent, dans la suite de ce chapitre, nous allons commencer par dériver une distribution purement audio adaptée à notre problème. Cette distribution doit être capable de modéliser fidèlement et efficacement le logarithme du module d'une variable aléatoire gaussienne circulaire complexe, ce qui doit donc s'appliquer au comportement des paramètres audio pour un son donné. Ensuite, nous proposerons une modélisation purement audio de la parole continue (*i.e.* comprenant plusieurs sons de parole différents) utilisant plusieurs noyaux de la distribution en question. Enfin, nous proposerons un modèle audiovisuel efficace fondé sur cette distribution que nous proposons pour les paramètres audio (3.1).

3.3.1 Modélisation statistique d'un seul son de parole

Pour des sections de parole quasi-stationnaires, les coefficients complexes $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T$ de la TFD peuvent être considérés comme ayant une dis-

tribution complexe gaussienne circulaire [89, 103] (certains auteurs [89] préfèrent employer le terme propre, *proper* en anglais, à la place de circulaire). Puisque les sons élémentaires sont considérés être à valeur moyenne nulle, les coefficients de leur TFD sont aussi à valeur moyenne nulle. De plus, il est classique que, pour des signaux stationnaires, les coefficients de la TFD soient décorrélés (et donc indépendants puisque gaussiens) ce qui conduit à une matrice de covariance Σ diagonale ([92] chapitre 4). Nous pouvons ainsi nous restreindre au cas d'une variable aléatoire scalaire (*i.e.* monodimensionnel) complexe gaussienne à valeur moyenne nulle sans avoir besoin d'étudier le cas vectoriel (*i.e.* multidimensionnel) comme développé dans la littérature [89, 103, 114].

Soit X une variable aléatoire complexe gaussienne circulaire à valeur moyenne nulle et de variance σ^2 : $X \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. La densité de probabilité de X est alors donnée par [89, 103]

$$p_X(x) = \left[\pi\sigma^2\right]^{-1} \exp\left[-\frac{|x|^2}{\sigma^2}\right]. \tag{3.6}$$

Cette équation suppose que les parties réelle et imaginaire de X sont décorrélées et de même variance égale à $\sigma^2/2$:

$$\begin{cases} \Re\{X\} & \sim & \mathcal{N}_{\mathbb{R}}\left(0, \frac{\sigma^2}{2}\right) \\ \Im\{X\} & \sim & \mathcal{N}_{\mathbb{R}}\left(0, \frac{\sigma^2}{2}\right) \end{cases}$$

où $\Re\{\cdot\}$ et $\Im\{\cdot\}$ sont respectivement les opérateurs partie réelle et partie imaginaire.

Une première solution non adaptée : un noyau gaussien complexe

Une première solution pour modéliser un son de parole est d'utiliser comme paramètres audio $\mathbf{a}(t)$ directement les coefficients complexes $\mathbf{S}(t)$ de la TFCT et ainsi de conserver un modèle à base de noyaux gaussiens, complexes dans ce cas $p(\mathbf{a}(t)) = p_G(\mathbf{a}(t)|0, \Sigma^A)$, où la loi de probabilité des paramètres audio est égale au produit des lois marginales à chaque fréquence f :

$$p_G(\mathbf{a}(t)|0, \Sigma^A) = \prod_{j=1}^{N_f} p_G(a(t, f_j)|0, \Sigma^A(f_j)).$$

Cependant, cette solution souffre d'un inconvénient majeur : les coefficients $S(t, f)$ de la TFCT sont mal conditionnés numériquement et plus particulièrement quand le nombre N_f de fréquences de calcul de la TFCT est grand. En effet, le maximum de $p_G(a(t, f)|0, \Sigma^A(f))$ varie en fonction de l'inverse de la racine carrée de la variance $\Sigma^A(f)$ donc $p_G(\mathbf{a}(t)|0, \Sigma^A)$ varie comme $(\prod_{j=1}^{N_f} \Sigma^A(f_j))^{-1}$. Or si ces variances sont grandes (*i.e.* très supérieures à $1/\pi$) et pour un grand nombre de dimensions (*i.e.* un grand nombre N_f de fréquences de calcul de la TFCT), ce produit tend à diminuer vers 0 causant des problèmes numériques à cause de la précision finie des calculs numériques[3]. Une telle modélisation est donc difficilement exploitable dans la pratique. C'est pourquoi nous choisissons bien définitivement pour paramètres audio $\mathbf{a}(t)$ le logarithme du module des coefficients de la TFCT et nous allons proposer une modélisation adaptée à ces coefficients.

[3]Pour plus de détails, se référer au paragraphe A.3 de l'annexe A.

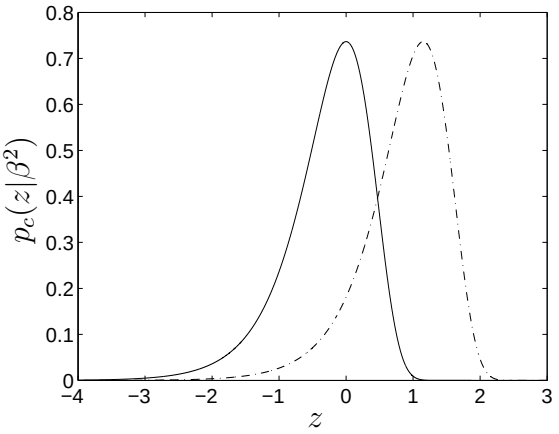


FIG. 3.4 – Densité de probabilité d'une loi LogRayleigh circulaire de paramètre de localisation $\beta^2 = 1$ (ligne continue) et $\beta^2 = 10$ (trait discontinu).

Une solution adaptée : la distribution de LogRayleigh

Distribution de LogRayleigh. Les coefficients $\mathbf{S}(t)$ suivant une loi normale centrée complexe circulaire, les paramètres audio $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ ne sont alors plus gaussiens. Ainsi, dans ce paragraphe, nous allons calculer la loi que suivent ces nouveaux paramètres audio [112].

Il est bien connu [92] que le module $Y = |X| = \sqrt{\Re\{X\}^2 + \Im\{X\}^2}$ de X (défini par l'équation (3.6)) est distribué suivant une loi de Rayleigh de paramètre $\sigma^2/2$: $Y \sim \text{Ray}(\sigma^2/2)$. La densité de probabilité d'une loi de Rayleigh de paramètre β^2 est donnée par

$$p_Y(y) = \begin{cases} \frac{y}{\beta^2} \exp\left(-\frac{y^2}{2\beta^2}\right) & \text{pour } y \geq 0, \\ 0 & \text{pour } y < 0. \end{cases} \tag{3.7}$$

Il est possible de montrer (*cf.* annexe A) que $Z = \ln Y$, défini comme le logarithme népérien de Y, suit une distribution de LogRayleigh (LR) de paramètre de localisation $\sigma^2/2$: $Z \sim \text{LogRay}(\sigma^2/2)$ dont l'expression est donnée à la définition 3.1 [112].

Définition 3.1 (Distribution de LogRayleigh)
Soit Z une variable aléatoire circulaire LogRayleigh (notée LR) de paramètre de localisation β^2, ce que nous noterons : $Z \sim \text{LogRay}(\beta^2)$. La fonction de densité de probabilité de cette variable est donnée par

$$\forall z \in \mathbb{R}, \quad p_Z(z) = p_c(z|\beta^2) = \frac{(e^z)^2}{\beta^2} \exp\left(-\frac{(e^z)^2}{2\beta^2}\right) \tag{3.8}$$

où $p_c(\cdot|\cdot)$ fait référence à la densité de probabilité circulaire. Cette distribution est représentée à la figure 3.4.

La distribution circulaire de LogRayleigh (LR) a une propriété intéressante : tous ses moments centrés d'ordre supérieur à un sont indépendants du paramètre de

localisation β^2. En effet, soient Z_1 et Z_2 deux variables LogRayleigh circulaires de paramètres de localisation respectifs β_1^2 et β_2^2, alors $p_{Z_1}(\cdot)$ et $p_{Z_2}(\cdot)$ vérifient

$$p_{Z_2}(z+\zeta) = p_{Z_1}(z), \quad \text{avec } \zeta = \ln \frac{\beta_2}{\beta_1}.$$

Ceci signifie que chaque distribution se déduit des autres par une translation dépendant des seuls paramètres de localisation respectifs (*cf.* figure 3.4).

Conséquence de la non-circularité. Jusqu'à maintenant, nous avons étudié le cas d'une variable aléatoire LogRayleigh circulaire. Cependant, dans certains cas, les coefficients de la TFD peuvent ne plus être circulaires. Ceci peut être dû à plusieurs causes et notamment au fait que le signal de parole est localement quasi-stationnaire mais pas strictement stationnaire. Dans ce paragraphe, nous étudions les conséquences de la non-circularité d'une variable aléatoire complexe Gaussienne à valeur moyenne nulle sur le logarithme du module de celle-ci.

Les moments d'ordre deux d'une variable aléatoire centrée complexe X sont la covariance $v_X \triangleq \mathrm{E}[xx^*]$ (où * signifie le complexe conjugué) et la pseudo-covariance $c_X \triangleq \mathrm{E}[xx]$ (*cf.* [89, 103] où ces moments sont définis dans le cas plus général multidimensionnel). Ainsi, dans le cas scalaire, nous avons

$$v_X = \sigma^2_{\Re(X)} + \sigma^2_{\Im(X)} \tag{3.9}$$

et

$$c_X = \sigma^2_{\Re(X)} - \sigma^2_{\Im(X)} + 2\jmath\,\rho\,\sigma_{\Re(X)}\,\sigma_{\Im(X)} \tag{3.10}$$

où $\jmath = \sqrt{-1}$, $\sigma^2_{\Re(X)}$ et $\sigma^2_{\Im(X)}$ sont respectivement les variances des parties réelle et imaginaire de X. ρ est le coefficient de corrélation entre les parties réelle et imaginaire de X, défini par

$$\rho = \frac{\mathrm{E}[\Re(X)\Im(X)]}{\sigma_{\Re(X)}\sigma_{\Im(X)}}.$$

Dans le cas circulaire, la pseudo-covariance est nulle ($c_X = 0$), ce qui signifie que l'on a à la fois $\sigma_{\Re(X)} = \sigma_{\Im(X)}$ et $\rho = 0$. Ainsi, la non-circularité peut se traduire soit par des autovariances différentes entre les parties réelle et imaginaire de X, soit par une corrélation entre celles-ci. Notons $\delta^2 = v_X$ dans un souci de simplicité et introduisons ϵ tel que

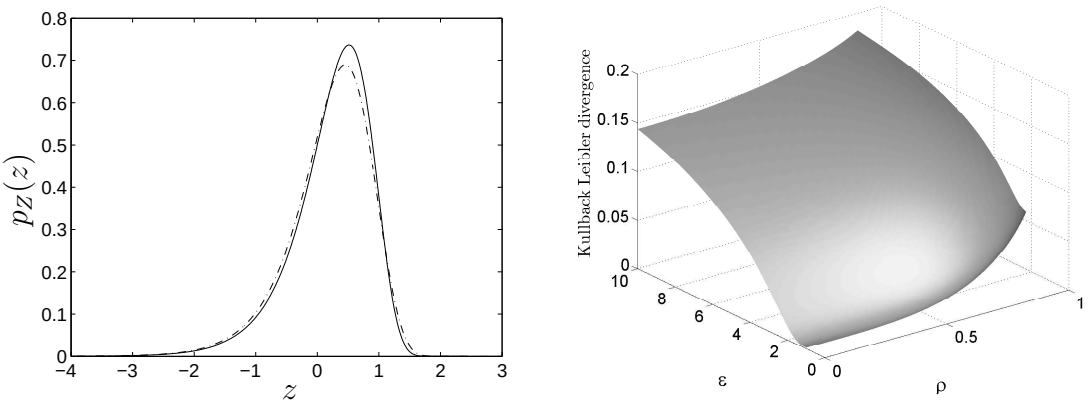
$$\sigma_{\Re(X)} = \epsilon\ \sigma_{\Im(X)}. \tag{3.11}$$

Dans ce cas, on peut montrer (*cf.* annexe A) que la densité de probabilité d'une variable aléatoire LR non-circulaire Z est égale à

$$p_Z(z) = p_c\left(z \,\middle|\, \frac{\delta^2}{2}\right) I(z, \delta^2, \rho, \epsilon) \tag{3.12}$$

avec $p_c(z|\beta^2)$ la densité de probabilité d'une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2 donné par l'équation (3.8) et

$$I(z, \delta^2, \rho, \epsilon) = \frac{\epsilon + 1/\epsilon}{2\sqrt{1-\rho^2}}\ \exp\left(-\frac{4\rho^2 + (\epsilon - 1/\epsilon)^2}{4(1-\rho^2)\delta^2}\,(e^z)^2\right) \times I_0\left(\left(\epsilon + \frac{1}{\epsilon}\right)\frac{\sqrt{(\epsilon - 1/\epsilon)^2 + 4\rho^2}}{4(1-\rho^2)\delta^2}(e^z)^2\right) \tag{3.13}$$



(a) (b)

FIG. 3.5 – Conséquences de la non-circularité. Fig. 3.5(a) : $\rho = -0.4$ et $\epsilon = 0.8$, les traits discontinus représentent la distribution LR non-circulaire (3.12) et la ligne continue la distribution LR circulaire optimale dont le paramètre de localisation est donné par (3.15). Fig.3.5(b) : divergence de Kullback-Leibler entre la densité de probabilité LR non-circulaire et LR circulaire optimale en fonction de ρ et ϵ.

où $I_0(\cdot)$ est la fonction de Bessel modifiée de première espèce :

$$I_0(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\{x \sin\theta\} d\theta.$$

Notons que, dans cette expression, la densité de probabilité d'une variable aléatoire LR non-circulaire est exprimée comme le produit d'une distribution LR circulaire dont le paramètre de localisation ne dépend que de la covariance δ^2 (et donc que d'ϵ) et d'un terme correcteur $I(\cdot,\cdot,\cdot,\cdot)$ qui dépend de la covariance δ^2 et de la pseudo-covariance c_X (et donc que d'ϵ et de ρ).

Bien que plus compliquée, cette distribution non circulaire est très proche d'une distribution circulaire lorsque ρ et $|1-\epsilon|$ sont petits devant 1 (*cf.* figure 3.5). Ainsi, dans ce cas, nous proposons de modéliser cette densité de probabilité LR non circulaire par une densité LR circulaire $p_c(z|\alpha)$ dont le paramètre de localisation α est estimé de telle sorte que ces deux densités de probabilité soient aussi proches que possible au sens d'un certain critère. Dans notre étude, le critère retenu est la divergence de Kullback-Leibler (d'autres critères tels que l'erreur quadratique moyenne peuvent être utilisés mais ils conduisent à des calculs plus compliqués du paramètre de localisation optimal sans résultat analytique). On a alors

$$\hat{\alpha} = \arg\min_{\alpha} g(\alpha) \tag{3.14}$$

où $g(\alpha) = KL[p_Z(\cdot)\|p_c(\cdot|\alpha)]$ est la divergence de Kullback-Leibler :

$$g(\alpha) = KL[p_Z(\cdot)\|p_c(\cdot|\alpha)] = \int_{-\infty}^{+\infty} p_Z(z) \ \ln\left(\frac{p_Z(z)}{p_c(z|\alpha)}\right) dz.$$

Dans la suite de notre étude, nous appellerons distribution LogRayleigh circulaire *optimale* la distribution LogRayleigh dépendant du paramètre de localisation optimal $\hat{\alpha}$.

Il est aisé de montrer que

$$\hat{\alpha} = \frac{1}{2}\int_{-\infty}^{+\infty} (e^z)^2 \; p_Z(z)\; dz$$

qui est égal à (*cf.* annexe A)

$$\hat{\alpha} = \frac{\delta^2}{2}. \tag{3.15}$$

On peut remarquer que puisque δ^2 est la somme des variances des parties réelle et imaginaire de X, ce paramètre de localisation optimal $\hat{\alpha}$ ne dépend que du paramètre ϵ (*cf.* équations (3.9) et (3.11)). Ainsi, ce paramètre de localisation $\hat{\alpha}$ définit une nouvelle variable aléatoire LR circulaire obtenue à partir d'une variable aléatoire LR non-circulaire en ne tenant pas compte du terme correcteur. Ceci implique que cette nouvelle variable aléatoire circulaire LR peut être vue comme le logarithme du module d'une variable aléatoire complexe gaussienne circulaire obtenue à partir d'une variable aléatoire complexe gaussienne non-circulaire en annulant la pseudo-covariance tout en gardant la même covariance.

Application de la loi LogRayleigh pour la modélisation statistique d'un seul son de parole

Rappelons que le vecteur audio (3.5) est défini par $\mathbf{a}(t) = \ln|\mathbf{S}(t)| \in \mathbb{R}^{N_f}$, où le vecteur $\mathbf{S}(t) = [S(t,f_1),\cdots,S(t,f_{N_f})]^T \in \mathbb{C}^{N_f}$ regroupe les coefficients complexes de la TFCT calculée sur 20ms. Les signaux de parole étant supposés quasi-stationnaires sur cette durée, les coefficients de la TCFT sont décorrélés à des fréquences différentes et donc indépendants car gaussiens (complexes), ce qui nous pousse à adopter des densités factorisables pour $\mathbf{a}(t)$. Un seul son de parole est ainsi modélisé par le produit des distributions optimales LogRayleigh marginales

$$p(\mathbf{a}(t)) = \prod_{j=1}^{N_f} p_{LR}\big(a(t,f_j)\,\big|\Gamma^A(f_j)\big)\,, \tag{3.16a}$$

où $\Gamma^A(f) = \Sigma^A(f)/2$ et $\Sigma^A(f)$ est la variance de $S(t,f)$. Dans un souci de simplicité, nous noterons $p_{LR}(\cdot|\Gamma)$ le produit des densités LR marginales $p_{LR}(\cdot|\gamma_i)$: $p_{LR}(\cdot|\Gamma) = \prod_{i=1}^{N} p_{LR}(\cdot|\gamma_i)$, la matrice de localisation Γ correspondante est alors diagonale : $\Gamma = \mathrm{diag}(\gamma_1,\cdots,\gamma_N)$. On peut ainsi réécrire (3.16a) de façon synthétique par

$$p(\mathbf{a}(t)) = p_{LR}\big(\mathbf{a}(t)\,\big|\Gamma^A\big) \tag{3.16b}$$

où la matrice de localisation Γ^A optimale est donnée par

$$\Gamma^A = \begin{pmatrix} \frac{\Sigma^A(f_1)}{2} & & 0 \\ & \ddots & \\ 0 & & \frac{\Sigma^A(f_{N_f})}{2} \end{pmatrix}.$$

Cette matrice est caractéristique de l'allure spectrale du son ainsi modélisé. Les coefficients diagonaux $\{\Sigma^A(f_1), \cdots, \Sigma^A(f_{N_f})\}$ de la matrice de covariance Σ^A des coefficients $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T$ de la TFCT peuvent être interprétés comme étant la densité spectrale de puissance du son considéré [16, 17]. Ainsi, la matrice de localisation Γ^A est également caractéristique de l'enveloppe spectrale du son modélisé par ce noyau.

3.3.2 Modélisation statistique de la parole continue

Dans ce paragraphe, nous caractérisons la modélisation des coefficients audio $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ pour de la parole continue. Pour cela, nous choisissons un modèle qui tient compte de la structure de la parole continue : celle-ci contenant plusieurs sons, nous adoptons un modèle multi-noyaux où chaque noyau doit idéalement modéliser un son. Nous avons identifié au paragraphe précédent que la distribution suivie par les coefficients spectraux $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ d'un son donné est une distribution de LogRayleigh optimale. Ainsi, nous proposons de modéliser les coefficients acoustiques de la parole continue par

$$p(\mathbf{a}(t)) = \sum_{i=1}^{N_A} \omega_i^A p_{LR}\big(\mathbf{a}(t) \,\big|\, \Gamma_i^A\big) \tag{3.17}$$

où ω_i^A et Γ_i^A sont respectivement le poids et la matrice de localisation du $i^{\text{ème}}$ noyau. La matrice de localisation Γ_i^A est alors caractéristique de l'allure spectrale du son modélisé par le $i^{\text{ème}}$ noyau, ω_i^A représentant sa probabilité *a priori* d'apparition. Ce modèle sera testé au paragraphe 3.5.1.

3.3.3 Modélisation audiovisuelle de la parole continue

Pour construire un modèle audiovisuel multi-noyaux adapté aux paramètres choisis et tel que chaque noyau audiovisuel modélise un son de parole donné, nous allons adopter la même démarche que celle qui nous a permis d'aboutir au modèle purement audio de la parole continu (3.17).

Rappelons tout d'abord que le vecteur vidéo $\mathbf{v}(t) = [v_w(t), v_h(t)]^T \in \mathbb{R}^2$ regroupe les largeur et hauteur intéro-labiales et que le vecteur audio est défini par $\mathbf{a}(t) = \ln|\mathbf{S}(t)| \in \mathbb{R}^{N_f}$, où $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T \in \mathbb{C}^{N_f}$ regroupe les coefficients complexes de la TFCT. Rappelons également que ces deux vecteurs sont extraits de façon synchrone. Comme précédemment, nous proposons, pour modéliser conjointement les paramètres vidéo $\mathbf{v}(t)$ et audio $\mathbf{a}(t)$, un modèle multi-noyaux où chacun d'eux doit modéliser un son particulier. Chaque noyau est choisi sous forme séparable : la densité des paramètres vidéo est choisie gaussienne et la densité des paramètres audio est une loi LogRayleigh optimale avec une matrice de localisation diagonale. On a alors

$$p_{AV}(\mathbf{a}(t), \mathbf{v}(t)) = \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_G\big(\mathbf{v}(t) \,\big|\, \mu_i^V, \Sigma_i^V\big)\; p_{LR}\big(\mathbf{a}(t) \,\big|\, \Gamma_i^A\big) \tag{3.18}$$

où la matrice diagonale de localisation Γ_i^A donnée par

$$\Gamma_i^A = \frac{\Sigma_i^A}{2}$$

est caractéristique de l'allure spectrale du $i^{\text{ème}}$ noyau. Σ_i^A est la diagonale de la matrice de covariance des coefficients $\mathbf{S}(t)$ de la TFCT.

Ce nouveau modèle audiovisuel, défini par l'équation (3.18) et paramétré par l'ensemble $\Theta = \{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Gamma_i^A\}_i$ qui devra être estimé, est construit en tenant compte des spécificités de la parole qui est composée de plusieurs sons. Chacun d'eux est ainsi supposé être modélisé individuellement par un noyau audiovisuel reliant d'une part la forme moyenne des lèvres définie par μ_i^V (Σ_i^V caractérisant la variabilité labiale) et d'autre part la densité spectrale moyenne du son prononcé dont le spectre est défini par la matrice diagonale de localisation Γ_i^A. Les poids ω_i^{AV} de ce nouveau modèle peuvent se voir comme les probabilités *a priori* de chacun des noyaux et donc comme la probabilité *a priori* d'apparition du son associé. Notons que bien que chaque noyau soit choisi avec une densité de probabilité audiovisuelle factorisable (*i.e.* égale au produit de la densité visuelle par la densité audio), il n'y a pas indépendance entre les données vidéo et audio puisque le modèle global (3.18) est la somme de plusieurs noyaux : le modèle (3.18) n'est donc pas factorisable.

3.3.4 Apprentissage des paramètres du modèle audiovisuel

De façon à estimer, à partir d'une base de donnée d'apprentissage, l'ensemble $\Theta = \{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Gamma_i^A\}_i$ regroupant les paramètres du modèle audiovisuel, nous proposons d'utiliser l'algorithme EM [46] (en anglais "Expectation-Maximisation") dans sa version pénalisée [91, 116]. Cet algorithme (*cf.* annexe B) permet d'estimer de façon itérative le jeu de paramètres Θ par la méthode du maximum de vraisemblance. Soit T le nombre de vecteurs audiovisuels, $\mathbf{a}(t)$ et $\mathbf{v}(t)$, dont nous disposons pour l'apprentissage. L'algorithme EM procède à chaque itération en deux étapes :

1. étape (E) : calcul de la probabilité *a posteriori* de chacun des noyaux connaissant les paramètres $\Theta^{(k)}$ à l'itération précédente k : $\forall i \in \{1, \cdots, N_{AV}\}$

$$p\Big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\Big) = \frac{\left(\omega_i^{AV}\right)^{(k)} \; p_G\Big(\mathbf{v}(t) \,\Big|\, \left(\mu_i^V\right)^{(k)}, \left(\Sigma_i^V\right)^{(k)}\Big) \; p_{LR}\Big(\mathbf{a}(t) \,\Big|\, \left(\Gamma_i^A\right)^{(k)}\Big)}{\sum_{j=1}^{N_{AV}} \left(\omega_j^{AV}\right)^{(k)} \; p_G\Big(\mathbf{v}(t) \,\Big|\, \left(\mu_j^V\right)^{(k)}, \left(\Sigma_j^V\right)^{(k)}\Big) \; p_{LR}\Big(\mathbf{a}(t) \,\Big|\, \left(\Gamma_j^A\right)^{(k)}\Big)}$$

   où $(\lambda)^{(k)}$ fait référence au paramètre λ à la $k^{\text{ème}}$ itération,

2. étape (M) : mise à jour des paramètres
   - poids audiovisuels ω_i^{AV}

$$\left(\omega_i^{AV}\right)^{(k+1)} = \frac{1}{T} \sum_{t=1}^{T} p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big)$$

   - paramètres vidéo

$$\left(\mu_i^V\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \big[\mathbf{v}(t) \; p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big)\big]}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)})}$$

et

$$\left(\Sigma_i^V\right)^{(k+1)} = \frac{\sum_{t=1}^{T}\left(\mathbf{v}(t)-\left(\mu_i^V\right)^{(k+1)}\right)\left(\mathbf{v}(t)-\left(\mu_i^V\right)^{(k+1)}\right)^T p\left(i|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)}\right)+2\alpha_i J_i}{\sum_{t=1}^{T} p(i|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)})+2\beta_i}.$$

où α_i, β_i et J_i sont les paramètres de pénalisation [91],

– paramètres audio, pour toutes les fréquences f_l

$$\left(\Gamma_i^A(f_l)\right)^{(k+1)} = \frac{\sum_{t=1}^{T}\left[\left(e^{a(t,f_l)}\right)^2 p\left(i|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)}\right)\right]}{2\ \sum_{t=1}^{T} p(i|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)})}.$$

La pénalisation de l'algorithme EM grâce aux paramètres α_i, β_i et J_i est nécessaire pour éviter la divergence de l'algorithme. En effet, la position fermée des lèvres a tendance à faire tendre vers 0 la variance du noyau correspondant. Les valeurs prises par la loi gaussienne ont alors tendance à tendre vers l'infini provoquant des problèmes numériques de calcul. Cette pénalisation garantit que les termes diagonaux de la variance ne tendent pas vers zéro.

3.4 Corpus

Pour tester notre modèle audiovisuel spécifique, nous avons utilisé deux corpus disponibles à l'ICP : l'un constitué de logatomes et l'autre de phrases.

Le premier corpus, que nous appellerons "corpus de logatomes", a été élaboré à l'ICP dans les années 90. Il s'agit d'un corpus audiovisuel monolocuteur constitué de logatomes français dénués de sens de la forme $[V_1-C-V_2-C-V_1]$ où V_1 et V_2 sont des voyelles identiques ou différentes parmi l'ensemble [a], [i], [y], [u] et C est une consonne parmi l'ensemble des plosives suivantes [p], [t], [k], [b], [d], [g] et [#], où [#] signifie l'absence de plosive. Cet ensemble de 112 séquences, représentant environ 50 secondes de parole, a été prononcé deux fois par le même locuteur masculin : la première série est utilisée pour l'apprentissage du modèle audiovisuel tandis que la seconde servira aux différents tests que ce soit sur le modèle lui-même ou pour la séparation ultérieurement. Ce corpus est intéressant car il regroupe dans un nombre réduit de logatomes les problèmes rencontrés par le traitement audiovisuel de la parole. Il contient d'une part des formes de lèvres similaires, telles que [y] et [u], associées à des sons différents et d'autre part des sons ayant certaines caractéristiques spectrales proches, tels que [y] et [i], mais ayant des formes de lèvres différentes. Les signaux visuels étant échantillonnés à 50Hz comme nous l'avons déjà mentionné, la longueur des trames audio est de 20ms. Ainsi, chacune des deux répétitions du corpus de logatomes contient environ 2500 trames audiovisuelles.

Le deuxième corpus, que nous appellerons "corpus de phrases", est composé de 107 phrases continues et phonétiquement équilibrées en français, prononcées par un même locuteur. Ce corpus représente environ 9300 trames, soit un peu plus de 3 minutes de parole. Il a pour but de faire apparaître de façon représentative les différentes occurrences des sons du français. Les phrases n'étant prononcées qu'une seule fois, nous avons divisé ce corpus en deux : les 80 premières phrases servant à l'apprentissage du modèle audiovisuel (soit environ 7200 trames correspondant à un

peu moins de 2 minutes et 30 secondes), les 27 dernières (soit environ 2100 trames correspondant à un peu plus de 20 secondes) étant réservées aux tests. La grande différence de ce corpus par rapport au précédent est une plus grande complexité, que ce soit par le nombre de sons présents ou par le fait que le corpus de test diffère du corpus d'apprentissage. En effet, avec ce corpus, les phrases présentes dans le corpus de test ne le sont pas dans le corpus d'apprentissage.

3.5 Expérimentations

Dans ce paragraphe, nous appliquons les résultats théoriques que nous venons de développer sur les deux corpus présentés ci-dessus. Nous présentons des résultats de modélisation du signal acoustique seul dans un premier temps, puis de modélisation du signal audiovisuel.

3.5.1 Modélisation audio

Dans ce paragraphe, nous comparons la modélisation du logarithme du module des coefficients de la TFD par des distributions LogRayleigh et par des distributions gaussiennes. Nous considérerons tout d'abord un son isolé de parole, puis de la parole continue. Pour ces expérimentations, puisque nous ne modélisons que la partie audio de la parole, les coefficients $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ sont calculés par TFCT sur 320 échantillons avec un recouvrement de 75% des trames (la contrainte de synchronie avec les paramètres vidéo étant levée).

Modélisation audio d'un seul son de la parole

Pour illustrer tout d'abord la modélisation audio d'un seul son de parole, toutes les trames identifiées comme une section de la voyelle [a] ont été extraites du corpus de logatome. Ainsi, environ 4 secondes ont été utilisées (représentant un total d'environ 800 vecteurs spectraux contenant les logarithmes des modules des coefficients de la TFD). On utilise ces 800 vecteurs pour calculer le paramètre de localisation optimal (3.15) (figure 3.6(a)) ainsi que ϵ et ρ pour chaque fréquence de calcul de la TFD (figure 3.6(b)). Le paramètre de localisation $\hat{\alpha}(f)$ peut être interprété comme la densité spectrale de puissance du son [17]. On constate sur la figure 3.6(c) que la densité LogRayleigh circulaire estimée suit relativement bien la distribution empirique (estimée par un histogramme) des coefficients audio de la voyelle. D'autre part, au moins deux noyaux gaussiens sont nécessaires pour modéliser de façon adéquate les mêmes données du fait de l'asymétrie de la distribution de LogRayleigh circulaire (figure 3.6(c)). Les paramètres des deux noyaux gaussiens ont été appris par l'algorithme EM sur les coefficients $a(t, f)$ où f est la $68^{\text{ème}}$ fréquence de calcul de la TFD : cela correspond donc à la densité de probabilité marginale à cette fréquence. De plus, modéliser de façon correcte la distribution LogRayleigh elle-même par un modèle multi-gaussien nécessite plus de noyaux gaussiens (typiquement quatre noyaux *cf.* figure 3.6(d)).

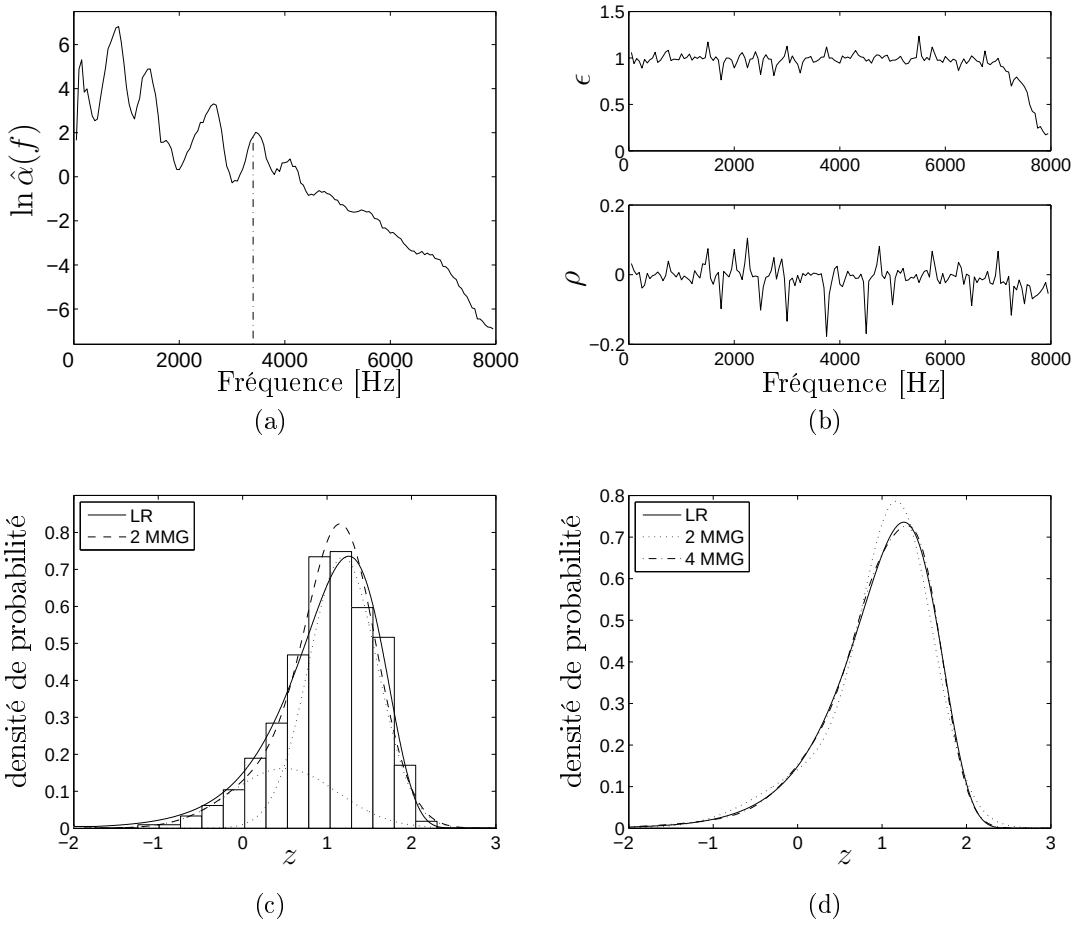


FIG. 3.6 – Modélisation de la voyelle [a]. Fig. 3.6(a) : logarithme du paramètre de localisation optimal en fonction de la fréquence, le trait discontinu correspond à la 68$^{\text{ème}}$ fréquence de calcul de la TFD. Fig. 3.6(b) : paramètres ϵ et ρ en fonction de la fréquence. Fig. 3.6(c) : histogramme à la 68$^{\text{ème}}$ fréquence de calcul de la TFD, la distribution LR circulaire optimale (ligne continue), l'approximation par 2 noyaux gaussiens (trait discontinu) et les deux noyaux correspondant (pointillé). Figure 3.6(d) : approximation de la loi LR par un modèle multi-gaussien (MMG).

Modélisation de la parole continue

Dans ce paragraphe, nous caractérisons la modélisation des coefficients audio $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ pour les corpus complets des logatomes d'une part et des phrases d'autre part. Nous modélisons donc la parole continue par un modèle multi-noyaux où chaque noyau modélise un son particulier. De plus, nous choisissons de comparer deux types de noyaux : les noyaux gaussiens et les noyaux LogRayleigh que nous avons proposés. Dans le cas de noyaux gaussiens, la densité de probabilité $p_{MMG}(\cdot)$ des paramètres $\mathbf{a}(t)$ s'écrit

$$p_{MMG}(\mathbf{a}(t)) = \sum_{i=1}^{N_A} \omega_i^A \, p_G\big(\mathbf{a}(t)|\mu_i^A, \Sigma_i^A\big) \tag{3.19}$$

où ω_i^A, μ_i^A et Σ_i^A sont les poids, les vecteurs des valeurs moyennes et les matrices de covariance du $i^{\text{ème}}$ noyau audio. Dans le cas de noyaux LogRayleigh optimaux, la densité de probabilité $p_{MMLR}(\cdot)$ des paramètres $\mathbf{a}(t)$ s'écrit

$$p_{MMLR}(\mathbf{a}(t)) = \sum_{i=1}^{N_A} \omega_i^A \, p_{LR}\big(\mathbf{a}(t)|\Gamma_i^A\big) \tag{3.20}$$

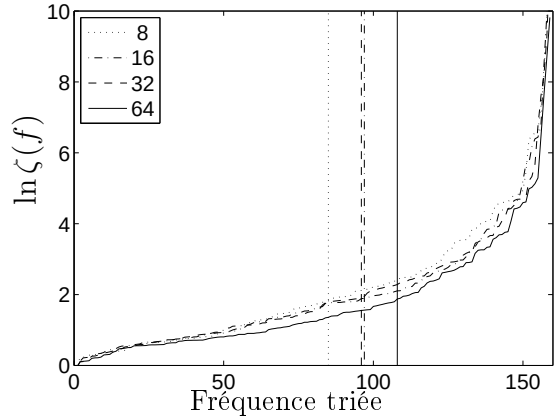
où ω_i^A et Γ_i^A sont les poids et matrice de localisation du $i^{\text{ème}}$ noyau audio. Pour ces deux types de modélisation, les paramètres $\{\omega_i^A, \mu_i^A, \Sigma_i^A\}_i$ et $\{\omega_i^A, \Gamma_i^A\}_i$ sont appris sur les corpus d'apprentissage par l'algorithme EM (*cf.* annexe B).

De façon à comparer l'adéquation des données au modèle, nous utilisons les corpus de test et nous calculons à chaque fréquence f, l'indice du test de Pearson [92]

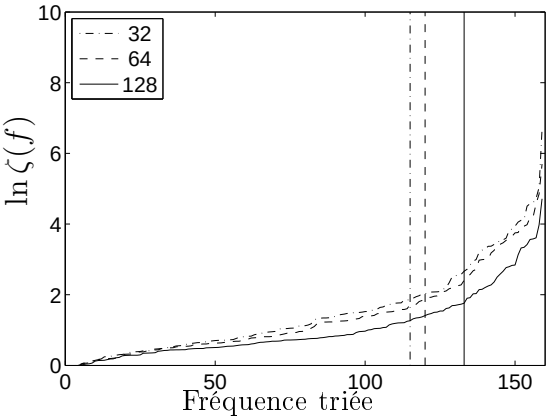
$$\zeta(f) = T \sum_{m=1}^{M} \frac{\big(\hat{p}_m(f) - p_m(f)\big)^2}{p_m(f)} \tag{3.21}$$

où T est le nombre de données à notre disposition, M le nombre de classes de l'histogramme utilisé pour calculer la distribution empirique, $\hat{p}_m(f)$ et $p_m(f)$ sont respectivement les probabilités empiriques et théoriques de la $i^{\text{ème}}$ classe à la fréquence f. Ainsi, $p_m(f) = \int_{x \in \mathcal{M}_m} p_{Mod}(x)dx$ où $\mathcal{M}_m$ est la $m^{\text{ème}}$ classe de l'histogramme et $p_{Mod}(x)$ la densité de probabilité marginale du modèle (3.19) ou (3.20) à la fréquence f. On compte ensuite le nombre de fréquences pour lesquelles (3.21) est plus petit qu'un certain seuil, par exemple le seuil de confiance à 5% défini par $\chi^2_{0.95}(M-1)$ si l'on utilise le test du χ^2 [92]. Pour montrer l'avantage du modèle multi-LogRayleigh (MMLR) par rapport au modèle multi-gaussien (MMG), nous comparons les résultats du test pour les deux modélisations.

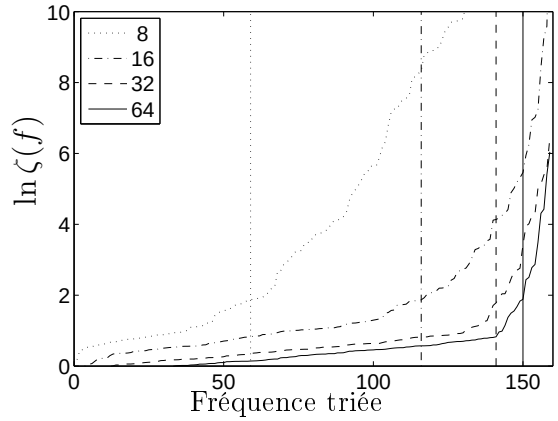
Dans le cas du corpus de logatomes (figure 3.7), on constate (figures 3.7(a) et 3.7(c)) que, lorsque le nombre de noyaux augmente, l'adéquation des données au modèle augmente plus vite dans le cas MMLR que dans le cas MMG. En effet, (3.21) décroît plus vite vers zéro pour le cas MMLR que MMG, quand le nombre de noyaux augmente. La figure 3.7(e) montre que, pour un même nombre de noyaux N_A, le modèle MMLR permet d'obtenir une meilleure adéquation des données au modèle que le modèle MMG, excepté lorsque N_A est inférieur à 16. En effet, puisque le corpus de logatomes contient environ 10 phonèmes, un minimum de 10 noyaux



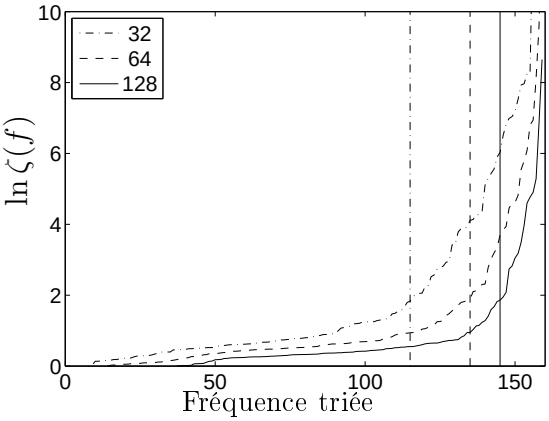
(a) MMG Logatomes



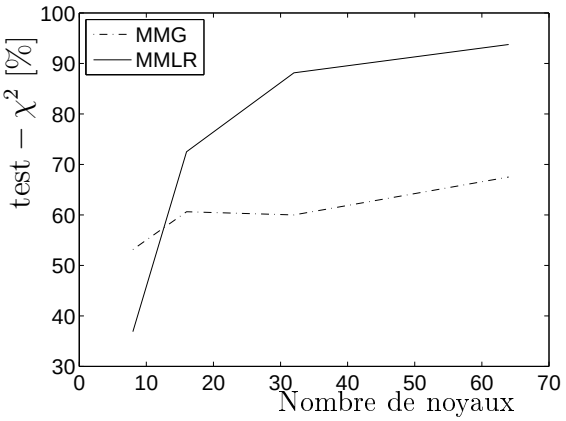
(b) MMG Phrases



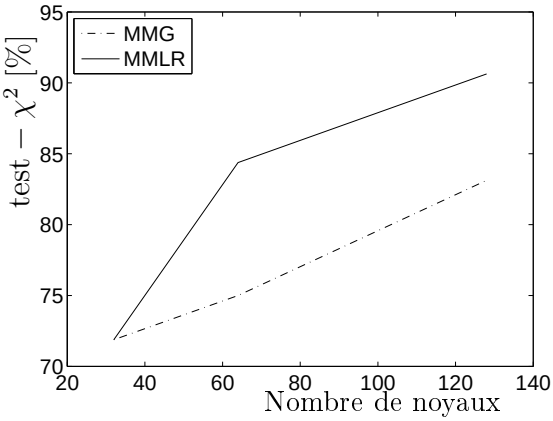
(c) MMLR Logatomes



(d) MMLR Phrases

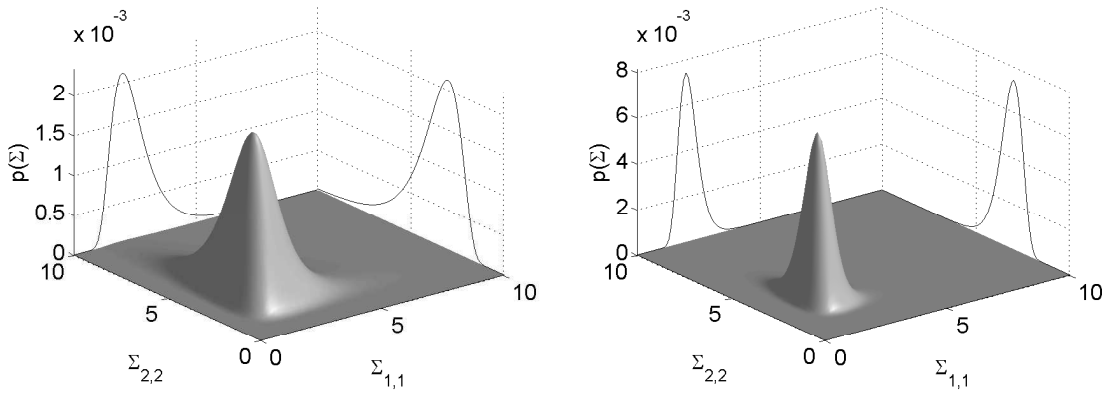


(e) Logatomes



(f) Phrases

FIG. 3.7 – Modélisation audio de la parole continue. Fig. 3.7(a), 3.7(c), Fig. 3.7(b) et 3.7(d) : logarithme de (3.21) (classé par ordre croissant) à chaque fréquence de calcul de la TFD pour les modèles MMG et MMLR (les légendes indiquent le nombre de noyaux) correspondant aux logatomes et aux phrases. Les droites verticales montrent le nombre de composantes qui vérifient le test du χ^2. Fig. 3.7(e) et 3.7(f) : pourcentage de composantes qui satisfont le test du χ^2 en fonction du nombre de noyaux pour le MMG (trait-point) et le MMLR (trait continu).



(a) $\alpha = 10,\ \beta = 10$ (b) $\alpha = 30,\ \beta = 30$

FIG. 3.8 – Loi de Wishart inverse $p(\Sigma)$ d'une matrice diagonale Σ. Les traits continus correspondent aux projections de la loi de Wishart inverse dans les plans $(\Sigma_{1,1}, 0)$ et $(0, \Sigma_{2,2})$. La densité de probabilité est donnée par $p(\Sigma) = p(\Sigma|\alpha, \beta, J) \propto \det(\Sigma^{-1})^{\beta} \exp[-\alpha \operatorname{Tr}(\Sigma^{-1} J)]$, avec dans notre exemple $J = \operatorname{diag}(2, 3)$ (*cf.* Annexe A).

est nécessaire pour le modèle MMLR. En fait, plus de 10 noyaux sont nécessaires puisque le corpus contient également des transitions entre phonèmes qu'il faut modéliser. D'autre part, pour une même fiabilité de modélisation, le modèle MMLR nécessite un nombre significativement moindre de noyaux que le modèle MMG. Par exemple, toujours pour les logatomes, les résultats obtenus avec 64 noyaux gaussiens sont les mêmes que ceux obtenus avec 16 noyaux LogRayleigh. Comme nous l'avons déjà mentionné, ceci est dû au fait que plusieurs noyaux gaussiens sont nécessaires pour modéliser correctement le même ensemble de données que modélise un unique noyau LogRayleigh.

Dans le cas du corpus de phrases (figures 3.7(b) et 3.7(d)), on peut faire le même constat : le modèle MMLR est plus performant que le modèle MMG pour modéliser la parole continue puisque (3.21) décroît plus vite vers zéro dans le cas MMLR que dans le cas MMG. De même, la figure 3.7(f) montre que, pour un même nombre de noyaux, le modèle MMLR fournit une meilleure adéquation des données au modèle. Cependant, obtenir les mêmes performances avec le corpus des phrases que celles obtenues avec le corpus des logatomes nécessite plus de noyaux. Ceci n'est pas étonnant puisque le corpus des phrases est plus complexe que le corpus des logatomes car il contient une plus grande diversité de sons, ce qui nécessite donc plus de noyaux pour le modéliser correctement.

3.5.2 Modélisation audiovisuelle

Dans ce paragraphe nous allons présenter les résultats de la modélisation audiovisuelle de la parole continue par le modèle (3.18) dont les paramètres ont été appris sur les corpus d'apprentissage par l'algorithme EM pénalisé comme expliqué au paragraphe 3.3.4.

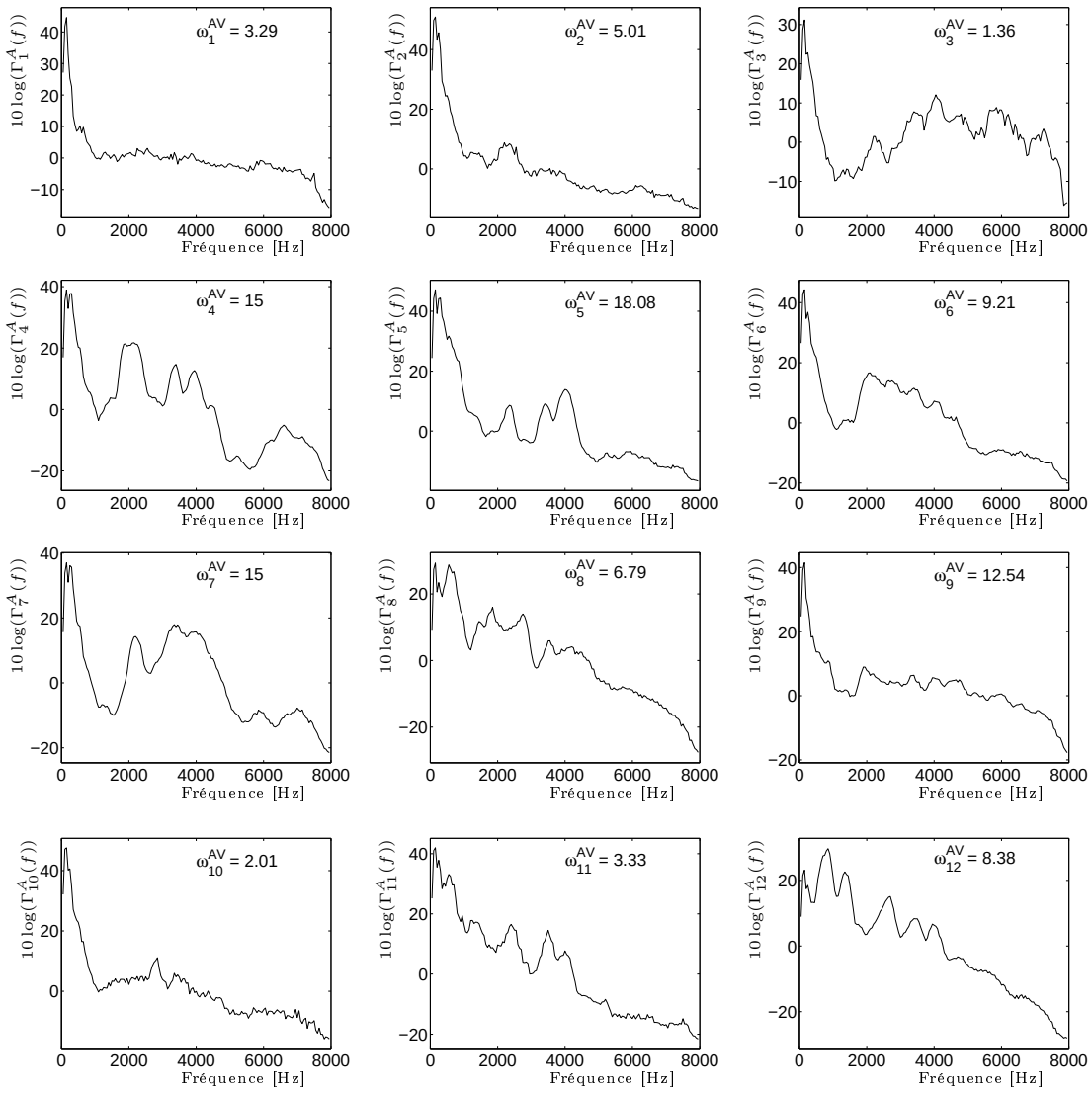
	[a]	[i]	[u]	[y]
F1 [Hz]	800	300	250	250
F2 [Hz]	1300	2200	(500)	1900
F3 [Hz]	2700	3300	2350	2200

TAB. 3.1 – Valeurs des formants de quatre voyelles du français mesurées sur les spectres issus de la modélisation. La valeur entre parenthèses a été estimée car elle n'est pas directement mesurable.

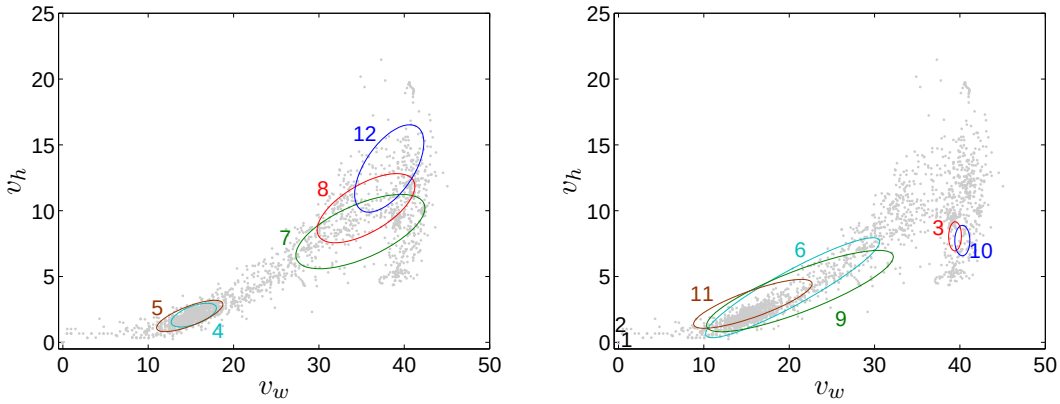
Lors de l'apprentissage pour le corpus de logatomes, nous avons choisi pour valeurs des paramètres de pénalisation $\alpha_i = 400$, $\beta_i = 400$ et $J_i = \mathrm{diag}(0.1, 0.1)$. Ceci permet de fortement pénaliser les valeurs prises par les matrices de covariance vidéo si celles-ci sont grandes devant celles de la matrice J comme illustré à la figure 3.8. En effet, d'une part, la loi de Wishart inverse (*cf.* Annexe B) est une loi unimodale dont le mode est obtenu en $\alpha J/\beta$ et la distribution de Wishart inverse devient de plus en plus piquée autour de son mode lorsque α et β augmente. Et d'autre part, à un son donné correspond une forme de lèvres particulière, ce qui nous pousse à vouloir choisir des noyaux bien localisé dans le plan (v_w, v_h). Ces deux considérations justifient notre choix des valeurs des paramètres de pénalisation.

La figure 3.9 illustre le résultat de l'apprentissage sur le corpus des logatomes du modèle (3.18) avec 12 noyaux. On constate en particulier que les noyaux 4, 5, 7 et 12 peuvent être associés respectivement aux voyelles [y], [u], [i] et [a]. En effet, on reconnaît d'une part facilement l'enveloppe spectrale typique de ces quatre voyelles et d'autre part la forme des lèvres associées. Dans l'espace vidéo (v_w, v_h), aux lèvres grandes ouvertes (noyau 12) correspond un [a], tandis que les noyaux 4 et 5 qui correspondent à des formes des lèvres arrondies avec une faible aire sont associés aux [y] et [u]. Des lèvres étirées (noyau 7) sont associées à un [i]. Certains noyaux modélisent des transitions entre phonèmes. Par exemple le noyau 8 modélise des transitions entre le [a] et les autres phonèmes. Nous avons remarqué que plus il y a de noyaux dans le modèle, plus il y a de noyaux qui modélisent les transitions. Le modèle multi-noyaux s'adapte ainsi à la richesse spectrale et labiale de la parole.

Il est très intéressant de constater que notre modélisation des coefficients audio $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$, qui utilise des densités LogRayleigh de matrice de localisation diagonale $\Gamma_i^A(f)$, permet de s'affranchir naturellement de l'influence de la fréquence fondamentale. Ainsi pour les voyelles, où cela se remarque le plus, les matrices de localisation des noyaux 4, 5, 7 et 12 donnent uniquement l'allure de l'enveloppe spectrale des sons correspondants. Il est même possible de faire une mesure des premiers formants de ces quatre voyelles comme reporté dans le tableau 3.1. Excepté pour le [u] où le deuxième formant est confondu dans le premier lobe du spectre, les mesures des trois premiers formants à partir des matrices de localisations des noyaux associés aux quatre voyelles donnent des valeurs cohérentes avec celles couramment reportées dans la littérature pour un homme [132, 133]. On retrouve également la complémentarité de la parole audiovisuelle évoquée au chapitre 1. En effet, les deux voyelles [u] et [y] sont proches dans l'espace vidéo mais ont des spectres différents. De la même façon les voyelles [u] et [i], qui ont des spectres assez proches, sont



(a) Matrices de localisation en dB



(b) Espace vidéo

FIG. 3.9 – Modèle audiovisuel des logatomes. Figure 3.9(a) : termes diagonaux des matrices de localisation $\Gamma_i^A(f)$ des 12 noyaux (en dB). Le poids w_i^{AV} du noyau correspondant est indiqué sur chaque graphe en %. Figure 3.9(b) : ellipses de confiance à 90% des paramètres vidéo des 12 noyaux (sur deux figures pour plus de clareté).

éloignées dans l'espace vidéo.

Pour le corpus de phrases on observe le même comportement bien que la modélisation complète du corpus soit plus compliquée. Ainsi, l'enveloppe spectrale des noyaux correspondant aux voyelles est plus grossière et la modélisation requiert plus de noyaux.

3.6 En résumé

Dans ce chapitre, nous avons caractérisé par un modèle multi-noyaux la distribution du logarithme des coefficients de la TFCT de la parole continue en exploitant sa structure non stationnaire. Chaque noyau a été construit de telle sorte qu'il modélise au mieux un unique son de la parole nous conduisant à dériver la distribution LogRayleigh plus adaptée et donc plus efficace que la distribution générale gaussienne. Comme l'ont montré nos expériences, pour caractériser la parole continue, une telle modélisation multi-LogRayleigh se montre plus appropriée qu'une modélisation multi-gaussienne car pour un même nombre de noyaux, elle requiert moins de paramètres puisque chaque noyau ne nécessite qu'un seul paramètre (de localisation) contre deux pour une gaussienne (moyenne et covariance) et de plus elle fournit une meilleure adéquation entre le modèle et les données. Finalement, nous avons étendu notre modélisation purement acoustique de façon à modéliser la parole audiovisuelle en proposant un modèle multi-noyaux où chaque noyau permet d'associer efficacement à une forme de lèvres donnée l'enveloppe spectrale du son correspondant.

Chapitre 4

La parole : un signal parcimonieux

La parole est un signal fortement non stationnaire : l'énergie du signal de parole sur une fenêtre d'analyse à court terme (*i.e.* de l'ordre de la dizaine de millisecondes ou moins) varie au cours du temps. En effet, la parole continue et spontanée est composée de différents sons enchaînés (co-articulés). A ces sons de parole se rajoutent d'autres sons produits par le locuteur tels que rires, bruits de respiration, grognements, *etc.* De plus, comme la parole spontanée comporte aussi des périodes de silence pendant lesquelles le locuteur ne produit aucun son, on peut dire que la parole est un signal parcimonieux. Dans certaines applications du traitement de la parole les périodes de silence sont exploitées, par exemple pour permettre l'apprentissage statistique du bruit environnant. Ainsi, en débruitage [82, 37], la détection des moments de silence permet d'améliorer les performances des algorithmes. Dans notre étude, la détection des moments de silence s'avérera cruciale pour les techniques de séparation de sources proposées dans la partie III. La détection de ces moments de silence sera soit intégrée dans des algorithmes de séparation de sources, soit à la base d'un nouveau principe d'extraction de source (*cf.* chapitre 6). Depuis le détecteur d'activité vocale (DAV) introduit par Freeman *et al.* en 1989 [53] fondé sur la recherche de déviation de caractéristiques spectrales du bruit et les travaux de Le Bouquin-Jeannès et Faucon [82] dont le DAV est fondé sur la fonction de cohérence, la détection d'activité vocale a été abordée d'un point de vue statistique par maximum de vraisemblance ou maximum *a posteriori* [121, 128, 55].

Dans ce chapitre, nous rappelons brièvement le principe d'un détecteur statistique d'activité vocale avant de l'étendre dans un cadre audiovisuel exploitant la bimodalité de la parole à travers le modèle statistique que nous avons proposé au chapitre 3. Nous introduirons ensuite un détecteur de silence (le silence étant défini ici comme la non-activité vocale d'un locuteur donné) exploitant uniquement la modalité visuelle. Enfin, nous présenterons les corpus utilisés avant de donner des résultats expérimentaux de notre modélisation.

4.1 Principe de la détection audio d'activité vocale

Le but de la détection d'activité vocale est d'attester de la présence ou non d'un signal de parole $s(t)$ à partir d'une observation bruitée $x(t) = s(t) + b(t)$, où $b(t)$ est le bruit (bruit environnant, signal de parole concurrent, *etc*). Pour cela, nous

allons recourir au test d'hypothèse. Soient H_0 l'hypothèse "le locuteur d'intérêt ne parle pas à l'instant t" (*i.e.* $s(t) = 0$) et H_1 l'hypothèse "le locuteur d'intérêt parle à l'instant t" (*i.e.* $s(t) \neq 0$). Dans ces conditions, les hypothèses s'écrivent

$$\begin{aligned} H_0 &: x(t) = b(t) \\ H_1 &: x(t) = s(t) + b(t) \end{aligned}$$

et la détection d'activité vocale est donnée par

$$\Pr[H_1|x(t)] \underset{H_0}{\overset{H_1}{\gtrless}} \Pr[H_0|x(t)] \tag{4.1}$$

où $\Pr[H_i|x(t)]$ est la probabilité *a posteriori* de l'hypothèse H_i sachant l'observation $x(t)$. Ce test peut se réécrire grâce à la formule de Bayes

$$p[x(t)|H_1]\Pr[H_1] \underset{H_0}{\overset{H_1}{\gtrless}} p[x(t)|H_0]\Pr[H_0] \tag{4.2}$$

en faisant intervenir les vraisemblances $p[x(t)|H_i]$ des hypothèses H_i et leur probabilité *a priori* $\Pr[H_i]$. En pratique, le calcul des vraisemblances est effectué dans le domaine fréquentiel, avec $\mathbf{X}(t) = [X(t, f_1), \cdots, X(t, f_{N_f})]^T$ le vecteur des coefficients de la transformée de Fourier à court terme (TFCT) à l'instant t de $x(t)$. Le test (4.2) devient alors

$$p[\mathbf{X}(t)|H_1]\Pr[H_1] \underset{H_0}{\overset{H_1}{\gtrless}} p[\mathbf{X}(t)|H_0]\Pr[H_0]. \tag{4.3}$$

Plusieurs modélisations des coefficients de la TFCT du signal et du bruit ont été présentées dans la littérature : gaussiennes [121] ou laplaciennes [55] par exemple conduisant à diverses expressions de l'équation (4.3) du détecteur d'activité vocale.

4.2 Détecteur audiovisuel d'activité vocale

Dans ce paragraphe, nous proposons d'utiliser le modèle à base de noyaux audiovisuels que nous avons construit au chapitre 3 pour caractériser les coefficients de la TFCT du signal $s(t)$ et le lien qui existe entre ces coefficients et les paramètres visuels du locuteur.

4.2.1 Principe de la détection audiovisuelle d'activité vocale

Supposons maintenant que nous ayons des observations audiovisuelles $s(t)$ et $\mathbf{v}(t) = [v_w(t), v_h(t)]^T$ du locuteur. Dans ces conditions, nous avons

– d'une part pour l'hypothèse H_0 "le locuteur ne parle pas"

$$H_0 : \begin{cases} \mathbf{X}(t) = \mathbf{B}(t) \\ \mathbf{v}(t) \end{cases} \tag{4.4}$$

avec

1. $\mathbf{B}(t) = [B(t, f_1), \cdots, B(t, f_{N_f})]^T$ est le vecteur des coefficients de la TFCT du bruit $b(t)$. Nous supposons ici le bruit stationnaire et gaussien centré, ainsi nous avons : $\mathbf{B}(t) \sim \mathcal{N}_{\mathbb{C}}(0, \Sigma_B)$,
2. le bruit $\mathbf{B}(t)$ et l'observation vidéo $\mathbf{v}(t)$ sont indépendants,

– et d'autre part pour l'hypothèse H_1 "le locuteur parle"[1]

$$H_1 \ : \ \begin{cases} \mathbf{X}(t) = \mathbf{S}(t) + \mathbf{B}(t) \\ \mathbf{v}(t) \end{cases} \tag{4.5}$$

avec

1. $\mathbf{S}(t) = [S(t, f_1), \cdots, S(t, f_{N_f})]^T$ est le vecteur des coefficients de la TFCT du signal de parole $s(t)$,
2. la parole $\mathbf{S}(t)$ et le bruit $\mathbf{B}(t)$ sont indépendants,
3. l'observation vidéo $\mathbf{v}(t)$ et le bruit $\mathbf{B}(t)$ sont indépendants,
4. le signal acoustique de parole $\mathbf{S}(t)$ et l'observation vidéo $\mathbf{v}(t)$ sont liés par le modèle audiovisuel(*cf.* chapitre 4)

$$p_{AV}(\mathbf{a}(t), \mathbf{v}(t)) = \sum_{i=1}^{N_{AV}} \omega_i^{AV} \, p_G\big(\mathbf{v}(t) \,|\mu_i^V, \Sigma_i^V\big) \; p_{LR}\big(\mathbf{a}(t) \,|\Gamma_i^A\big)$$

où $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$.

Le détecteur d'activité vocale fonctionne donc sur le principe de la comparaison des deux hypothèses H_0 et H_1 :

$$\Pr[H_1 \,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \underset{H_0}{\overset{H_1}{\gtrless}} \Pr[H_0 \,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \tag{4.6}$$

où nous utilisons le logarithme du module du vecteur des coefficients de la TFCT du signal acoustique $x(t)$ pour les raisons de conditionnement numérique que nous avons abordées au paragraphe 3.3.1. Ainsi, le détecteur audiovisuel d'activité vocale fait intervenir les probabilités *a posteriori* des deux hypothèses H_0 et H_1 connaissant les observations acoustiques et visuelles. Nous allons maintenant développer l'expression de ces probabilités.

Expression de $\Pr[H_0 \,|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$

D'après la règle de Bayes, la probabilité *a posteriori* de l'hypothèse H_0 s'écrit

$$\Pr[H_0 \,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto p[\ln|\mathbf{X}(t)|, \mathbf{v}(t)|H_0] \Pr[H_0].$$

Or les coefficients audio $\ln|\mathbf{X}(t)| = \ln|\mathbf{B}(t)|$ et visuels $\mathbf{v}(t)$ sont indépendants sous l'hypothèse H_0 puisque le bruit et l'observation vidéo sont indépendants. Ainsi, nous pouvons factoriser la vraisemblance

$$\Pr[H_0 \,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto p[\mathbf{v}(t)|H_0] \; p\left[\ln|\mathbf{X}(t)| \,\middle|H_0\right] \; \Pr[H_0]$$

[1]Nous verrons au paragraphe 4.2.2 comment modifier cette expression pour tenir compte du facteur d'amplitude entrant dans la modélisation audiovisuelle.

et finalement

$$\Pr[H_0\,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto p[\mathbf{v}(t)|H_0]\; p_{LR}\left(\ln|\mathbf{X}(t)|\,\middle|\,\frac{\Sigma_B}{2}\right)\; \Pr[H_0] \tag{4.7}$$

où $p[\mathbf{v}(t)|H_0]$ est le modèle statistique des lèvres dans le silence que nous choisissons sous forme d'un modèle multi-gaussien :

$$p[\mathbf{v}(t)|H_0] = \sum_{i=1}^{N_{H_0}} \omega_i^{H_0}\, p_G\big(\mathbf{v}(t)|\mu_i^{H_0}, \Sigma_i^{H_0}\big)\,. \tag{4.8}$$

$\omega_i^{H_0}$, $\mu_i^{H_0}$ et $\Sigma_i^{H_0}$ sont les paramètres du modèle des lèvres dans le silence.

Expression de $\Pr[H_1\,|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$

D'après la règle de Bayes, la probabilité *a posteriori* de l'hypothèse H_1 s'écrit

$$\Pr[H_1\,|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto p[\ln|\mathbf{X}(t)|, \mathbf{v}(t)|H_1]\,\Pr[H_1].$$

Or les coefficients audio $\ln|\mathbf{X}(t)|$ sont maintenant issus du mélange entre le bruit $\mathbf{B}(t)$ et le signal de parole $\mathbf{S}(t)$: $\ln|\mathbf{X}(t)| = \ln|\mathbf{S}(t)+\mathbf{B}(t)|$. De plus les coefficients de parole $\mathbf{a}(t) = \ln|\mathbf{S}(t)|$ et visuels $\mathbf{v}(t)$ sont liés par le modèle audiovisuel du chapitre 3. Ainsi sous l'hypothèse H_1, $\mathbf{X}(t)$ est la somme de deux variables aléatoires indépendantes $\mathbf{S}(t)$ et $\mathbf{B}(t)$. Or le bruit suit une loi normale centrée de matrice de covariance Σ_B et pour chaque noyau i $\mathbf{S}(t)$ suit une loi normale centrée de matrice de covariance $2\Gamma_i^A$. Donc, pour chaque noyau i, $\mathbf{X}(t)$ est une variable aléatoire gaussienne centrée de matrice de covariance $2\Gamma_i^A + \Sigma_B$ et finalement, le logarithme du module de $\mathbf{X}(t)$ suit une loi de LogRayleigh de matrice de localisation $\Gamma_i^A + \Sigma_B/2$. Nous pouvons en déduire alors que la relation entre les observations acoustique $\ln|\mathbf{X}(t)|$ et visuelle $\mathbf{v}(t)$ est donnée par

$$p[\ln|\mathbf{X}(t)|, \mathbf{v}(t)|H_1] = \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_G\big(\mathbf{v}(t)\,\big|\mu_i^V, \Sigma_i^V\big)\; p_{LR}\left(\ln|\mathbf{X}(t)|\,\middle|\,\Gamma_i^A + \frac{\Sigma_B}{2}\right). \tag{4.9}$$

4.2.2 Facteur d'amplitude

Modèle normalisé

Nous soulevons maintenant une difficulté de la modélisation audiovisuelle. En effet, le modèle audiovisuel que nous avons appris porte sur des coefficients spectraux normalisés : les paramètres audio exploités pour l'apprentissage du modèle (3.18) sont normalisés car de puissance unitaire. Or ici les coefficients $\mathbf{X}(t)$ du signal observé sont le résultat de la contribution du signal sans bruit $\mathbf{S}(t)$ et de celle du bruit $\mathbf{B}(t)$. Il n'est donc pas possible de les normaliser de telle sorte que la puissance de la contribution du signal $\mathbf{S}(t)$ à $\mathbf{X}(t)$ soit unitaire. Nous devons donc reparamétriser notre problème. Ainsi, nous introduisons maintenant le facteur d'amplitude $\sqrt{\alpha(t)}$ tel que $\sqrt{\alpha(t)}\,\mathbf{S}'(t) = \mathbf{S}(t)$ où $\mathbf{S}'(t)$ est le signal normalisé (*i.e.* de puissance unitaire)

sur lequel porte le modèle (3.18). Nous pouvons ainsi réécrire l'équation (4.5) de la façon suivante ((4.4) est inchangée)

$$H_1 \ : \ \begin{cases} \mathbf{X}(t) = \sqrt{\alpha(t)}\,\mathbf{S}'(t) + \mathbf{B}(t) \\ \mathbf{v}(t) \end{cases} \tag{4.10}$$

où $\alpha(t)$ est le facteur d'amplitude non négatif qu'il faudra estimer car le modèle audiovisuel que nous avons appris porte sur $\mathbf{a}(t) = \ln|\mathbf{S}'(t)|$ et $\mathbf{v}(t)$:

$$p_{AV}(\mathbf{a}(t), \mathbf{v}(t)) = \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_G\big(\mathbf{v}(t)\,\big|\mu_i^V, \Sigma_i^V\big)\; p_{LR}\big(\mathbf{a}(t)\,\big|\Gamma_i^A\big)\,.$$

Dans ces conditions la probabilité *a posteriori* de l'hypothèse H_1 devient

$$\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_G\big(\mathbf{v}(t)\,\big|\mu_i^V, \Sigma_i^V\big)\; p_{LR}\left(\ln|\mathbf{X}(t)|\,\middle|\alpha(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\right) \Pr[H_1] \tag{4.11}$$

qui fait donc intervenir le facteur d'amplitude inconnu $\alpha(t)$ qu'il est alors nécessaire d'estimer.

Estimation du facteur d'amplitude

Pour estimer le facteur d'amplitude $\alpha(t)$, nous proposons d'utiliser le principe du maximum de vraisemblance :

$$\hat{\alpha}(t) = \arg\max_{\alpha(t)} p[\mathbf{X}(t)|\alpha(t)] \tag{4.12}$$

où

$$p[\mathbf{X}(t)|\alpha(t)] = \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_{LR}\left(\ln|\mathbf{X}(t)|\,\middle|\alpha(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\right).$$

Recourir à cette expression pour l'estimation, revient à considérer que tous les noyaux ont généré le son et à chercher le facteur d'amplitude qui maximise la vraisemblance. Or cette expression n'est ni facile à optimiser, ni en adéquation avec notre modèle qui suppose qu'un son est généré par un seul noyau. Ainsi, nous proposons de remplacer l'estimation du facteur d'amplitude unique pour tous les noyaux par l'estimation de facteurs d'amplitude potentiellement différents pour chacun des noyaux

$$\hat{\alpha}_i(t) = \arg\max_{\alpha_i(t)} p_{LR}\left(\ln|\mathbf{X}(t)|\,\middle|\alpha_i(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\right). \tag{4.13}$$

Ainsi, $\hat{\alpha}_i(t)$ doit annuler la dérivée du logarithme de $p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\alpha_i(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\big)$ par rapport à $\alpha_i(t)$ sous contrainte de positivité de $\alpha_i(t)$:

$$\begin{aligned} 0 &= \frac{\partial p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\alpha_i(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\big)}{\partial \alpha_i(t)} \\ &= \sum_{j=1}^{N_f} \frac{\Gamma_i^A(f_j)}{\left[\frac{\Sigma_B(f_j)}{2} + \alpha_i(t)\Gamma_i^A(f_j)\right]^2} \left[\frac{|X(t,f_j)|^2}{2} - \frac{\Sigma_B(f)}{2} - \alpha_i(t)\,\Gamma_i^A(f_j)\right] \end{aligned} \tag{4.14}$$

avec $\alpha_i(t) \geq 0$. Pour résoudre cette équation sous contrainte, nous proposons de recourir à un algorithme itératif inspiré de [16, 17]. A l'itération $k+1$, nous supposons que le dénominateur de (4.14) ne dépend pas de $\big(\alpha_i(t)\big)^{(k+1)}$ mais de $\big(\alpha_i(t)\big)^{(k)}$, ainsi nous avons :

$$\big(\hat{\alpha}_i(t)\big)^{(k+1)} = \frac{\sum_{j=1}^{N_f} \left| \frac{|X(t,f_j)|^2}{2} - \frac{\Sigma_B(f_j)}{2} \right| \frac{\Gamma_i^A(f_j)}{\left[\frac{\Sigma_B(f_j)}{2} + (\hat{\alpha}_i(t))^{(k)} \Gamma_i^A(f_j)\right]^2}}{\sum_{j=1}^{N_f} \frac{\big(\Gamma_i^A(f_j)\big)^2}{\left[\frac{\Sigma_B(f_j)}{2} + (\hat{\alpha}_i(t))^{(k)} \Gamma_i^A(f_j)\right]^2}} \tag{4.15a}$$

où la valeur absolue assure la positivité de $\alpha_i(t)$. Pour l'initialisation, nous proposons de choisir par exemple

$$\big(\hat{\alpha}_i(t)\big)^{(0)} = \frac{\sum_{j=1}^{N_f} \left| \frac{|X(t,f_j)|}{2} - \frac{\Sigma_B^2(f)}{2} \right| \Gamma_i^A(f_j)}{\sum_{j=1}^{N_f} \big(\Gamma_i^A(f_j)\big)^2}. \tag{4.15b}$$

Influence du facteur d'amplitude sur la décision

Le test d'hypothèse (4.6) est donné par

$$\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \underset{H_0}{\overset{H_1}{\gtrless}} \Pr[H_0|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$$

où la probabilité *a posteriori* $\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$ est maintenant définie par

$$\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto \sum_{i=1}^{N_{AV}} \left[\omega_i^{AV} p_G\big(\mathbf{v}(t) \,|\mu_i^V, \Sigma_i^V\big)\, p_{LR}\left(\ln|\mathbf{X}(t)| \,\middle|\, \hat{\alpha}_i(t)\, \Gamma_i^A + \frac{\Sigma_B}{2} \right) \right] \Pr[H_1] \tag{4.16}$$

et

$$\Pr[H_0|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto \left[\sum_{i=1}^{N_{H_0}} \omega_i^{H_0} p_G\big(\mathbf{v}(t)|\mu_i^{H_0}, \Sigma_i^{H_0}\big) \right] p_{LR}\left(\ln|\mathbf{X}(t)| \,\middle|\, \frac{\Sigma_B}{2} \right) \Pr[H_0]. \tag{4.17}$$

Dans le cas d'une trame de silence à l'instant t, si l'on estime parfaitement les facteurs d'amplitude, *i.e.* $\forall i,\ \hat{\alpha}_i(t) = 0$, alors l'expression de $\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$ se réduit à

$$\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|] \propto \left[\sum_{i=1}^{N_{AV}} \omega_i^{AV} p_G\big(\mathbf{v}(t) \,|\mu_i^V, \Sigma_i^V\big) \right] p_{LR}\left(\ln|\mathbf{X}(t)| \,\middle|\, \frac{\Sigma_B}{2} \right) \Pr[H_1].$$

En posant $p[\mathbf{v}(t)|H_1] = \sum_{i=1}^{N_{AV}} \omega_i^{AV} p_G\big(\mathbf{v}(t) \,|\mu_i^V, \Sigma_i^V\big)$, l'équation du test d'hypothèse devient

$$p[\mathbf{v}(t)|H_1] \Pr[H_1] \underset{H_0}{\overset{H_1}{\gtrless}} p[\mathbf{v}(t)|H_0] \Pr[H_0].$$

Ainsi, dans ce cas toute la décision de détection repose entièrement sur la partie vidéo du test : il est alors important d'avoir un bon modèle des lèvres du silence. Cependant, l'estimation des facteurs d'amplitude $\alpha_i(t)$ ne sera jamais parfaite et par construction on a

$$\forall i, \quad p_{LR}\left(\ln|\mathbf{X}(t)| \,\middle|\, \hat{\alpha}_i(t)\Gamma_i^A + \frac{\Sigma_B}{2}\right) \geq p_{LR}\left(\ln|\mathbf{X}(t)| \,\middle|\, \frac{\Sigma_B}{2}\right)$$

ce qui implique des erreurs de type non détection de silence plus importantes.

4.2.3 Mise à jour des paramètres du silence

L'équation de décision de notre détecteur audiovisuel d'activité vocale (4.6) fait intervenir la variance Σ_B des coefficients complexes de la TFCT du bruit, ainsi que les paramètres $\{\omega_i^{H_0}, \mu_i^{H_0}, \Sigma_i^{H_0}\}_i$ du modèle des lèvres dans le silence (4.8). De façon à pouvoir s'adapter à d'éventuels changements de ces caractéristiques nous proposons dans ce paragraphe une amélioration possible du détecteur audiovisuel d'activité vocal en proposant des équations de mises à jour de ces paramètres.

Variance du bruit

Il est classique pour mettre à jour les variances $\Sigma_B(f)$ de recourir à des équations avec facteur d'oubli exponentiel [51, 50]. A l'indice temporel $k+1$ des trames (*i.e.* indices temporels des transformées de Fourier à court terme) on a ainsi

$$(\Sigma_B(f))^{(k+1)} = \begin{cases} \gamma\,(\Sigma_B(f))^{(k)} + (1-\gamma)\,|X(k+1,f)|^2 & \text{sous } H_0 \\ (\Sigma_B(f))^{(k)} & \text{sous } H_1 \end{cases}$$

où $0 < \gamma < 1$ est un paramètre de lissage (ou facteur d'oubli) exponentiel. Cependant cela nécessite à chaque itération k de tester si l'on a détecté du silence H_0 ou de la parole H_1. Pou simplifier, il est possible d'adopter une seule équation pour les deux décisions :

$$(\Sigma_B(f))^{(k+1)} = \gamma\,(\Sigma_B(f))^{(k)} + (1-\gamma)\,E\left[|B(k+1,f)|^2 \,\middle|\, \mathbf{v}(k+1), \ln|\mathbf{X}(k+1)|\right]$$

avec

$$E\left[|B(k,f)|^2 \,\middle|\, \mathbf{v}(k), \ln|\mathbf{X}(k)|\right] = |X(k,f)|^2 \Pr\left[H_0 \middle| \mathbf{v}(k), \ln|\mathbf{X}(k)|\right] + (\Sigma_B(f))^{(k-1)} \Pr\left[H_1|\mathbf{v}(k), \ln|\mathbf{X}(k)|\right].$$

Cette expression peut s'interpréter ainsi : si l'on détecte de la parole, on conserve la valeur précédente de la variance du bruit, tandis que si l'on détecte du silence, on exploite l'observation pour mettre à jour la variance du bruit. Ainsi, nous avons

$$(\Sigma_B(f))^{(k+1)} = \tilde{\gamma}_{k+1}\,(\Sigma_B(f))^{(k)} + (1-\tilde{\gamma}_{k+1})\,|X(k+1,f)|^2 \qquad (4.18)$$

où $\tilde{\gamma}_{k+1} = \gamma + (1-\gamma)\Pr[H_1|\mathbf{v}(k+1), \ln|\mathbf{X}(k+1)|]$. Cette dernière expression de mise à jour fait intervenir un facteur d'oubli exponentiel $\tilde{\gamma}_{k+1}$ dépendant de la probabilité *a posteriori* d'avoir de la parole. Elle permet ainsi de regrouper en une seule équation les deux cas survenant à chaque hypothèse (silence ou parole).

Paramètres du modèle des lèvres dans le silence

Pour permettre l'adaptation des paramètres $\Theta_{H_0} = \{\omega_i^{H_0}, \mu_i^{H_0}, \Sigma_i^{H_0}\}_i$ du modèle des lèvres dans le silence (4.8), nous proposons l'algorithme suivant en deux étapes inspiré de l'algorithme EM pénalisé [91, 116].

La première étape consiste à calculer les probabilités *a posteriori* des noyaux pour les observations vidéo $\mathbf{v}(t)$, $\forall\ 1 < t < k+1$:

$$\forall i, \quad p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big] = \frac{\left(\omega_i^{H_0}\right)^{(k)}\ p\Big[\mathbf{v}(t) \,\Big|\, \left(\mu_i^{H_0}\right)^{(k)}, \left(\Sigma_i^{H_0}\right)^{(k)}\Big]}{\sum_{j=1}^{N_{H_0}} \left(\omega_j^{H_0}\right)^{(k)}\ p\Big[\mathbf{v}(t) \,\Big|\, \left(\mu_j^{H_0}\right)^{(k)}, \left(\Sigma_j^{H_0}\right)^{(k)}\Big]}.$$

La seconde étape consiste à mettre à jour l'ensemble des paramètres Θ_{H_0}

$$\left(\omega_i^{H_0}\right)^{(k+1)} = \frac{1}{k+1}\sum_{t=1}^{k+1} p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big]\ \Pr[H_0|k+1] + \left(\omega_i^{H_0}\right)^{(k)}\ (1 - \Pr[H_0|k+1]) \quad (4.19a)$$

où l'on note de façon synthétique $\Pr[H_0|k+1]$ la probabilité *a posteriori* que la trame à l'instant $k+1$ soit du silence : $\Pr[H_0|k+1] = \Pr[H_0|\mathbf{v}(k+1), \ln|\mathbf{X}(k+1)|]$. Pour les vecteurs des valeurs moyennes, on propose

$$\left(\mu_i^{H_0}\right)^{(k+1)} = \frac{\sum_{t=1}^{k+1} \mathbf{V}(t) p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big] + (k+1)\eta\left(\mu_i^{H_0}\right)^{(0)}}{\sum_{t=1}^{k+1} p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big] + (k+1)\eta}\ \Pr[H_0|k+1] + \left(\mu_i^{H_0}\right)^{(k)}\ (1 - \Pr[H_0|k+1]) \quad (4.19b)$$

η étant un paramètre de pénalisation et $\left(\mu_i^{H_0}\right)^{(0)}$ le vecteur initial des valeurs moyennes. Pour les matrices de covariances, on propose

$$\left(\Sigma_i^{H_0}\right)^{(k+1)} = \left(\frac{\sum_{t=1}^{k+1}\left(\mathbf{v}(t) - \left(\mu_i^{H_0}\right)^{(k+1)}\right)\left(\mathbf{v}(t) - \left(\mu_i^{H_0}\right)^{(k+1)}\right)^T p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big]}{\sum_{t=1}^{k+1} p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big] + 2\beta + (k+1)\eta} + \frac{2\alpha + (k+1)\eta\left(\Sigma_i^{H_0}\right)^{(0)}}{\sum_{t=1}^{k+1} p\Big[i \,\Big|\, \mathbf{v}(t), (\Theta_{H_0})^{(k)}\Big] + 2\beta + (k+1)\eta}\right)\ \Pr[H_0|k+1] + \left(\Sigma_i^{H_0}\right)^{(k)}\ \left(1 - \Pr[H_0|k+1]\right) \quad (4.19c)$$

où $\left(\Sigma_i^{H_0}\right)^{(0)}$ est la variance initiale du $i^{\text{ème}}$ noyau. De façon à éviter les problèmes de temps de calcul, nous proposons de limiter les sommations des équations précédentes aux N dernières trames détectées comme étant du silence. Il est de plus intéressant de noter que, dans les équations (4.19b) et (4.19c), les termes de pénalisation $(k+1)\eta$ s'apparentent à des coefficients de raideur de force de rappel élastique vers les noyaux originaux. Ceci permet aux noyaux de revenir vers leur position initiale lorsque peu de vecteurs $\mathbf{v}(t)$ ont une grande probabilité *a posteriori* d'avoir été générés par ceux-ci.

4.2.4 Intégration temporelle

De façon à augmenter les performances de notre détection d'activité vocale, nous proposons d'intégrer temporellement de deux manières les équations permettant la décision du détecteur audiovisuel d'activité vocale.

La première intégration que nous utilisons consiste à intégrer les vraisemblances généralisées $L_{H_0}(t)$ et $L_{H_1}(t)$ (produit des vraisemblances par les probabilités *a priori*) des deux hypothèses [34] :

$$L_{H_1}(t) = (L_{H_1}(t-1))^{\kappa} \ (p[\ln|\mathbf{X}(t)|, \mathbf{v}(t)|H_1] \ \Pr[H_1])^{1-\kappa} \quad (4.20a)$$

$$L_{H_0}(t) = (L_{H_0}(t-1))^{\kappa} \ (p[\ln|\mathbf{X}(t)|, \mathbf{v}(t)|H_0] \ \Pr[H_0])^{1-\kappa} \quad (4.20b)$$

où κ est un facteur d'oubli. Ainsi, l'équation de décision du détecteur audiovisuelle d'activité vocale (4.6) devient

$$L_{H_1}(t) \underset{H_0}{\overset{H_1}{\gtrless}} L_{H_0}(t). \quad (4.21)$$

Nous proposons de plus une variante à cette intégration : elle consiste à décider que la trame à l'instant t est du silence si le produit des probabilités *a posteriori* des N trames $\{t, \cdots, t-N+1\}$ de l'hypothèse H_0 est supérieur à $1/2$, c'est-à-dire à la probabilité de l'ensemble des autres possibilités :

$$\Pr[H_0|\ln|\mathbf{X}(t)|, \mathbf{v}(t)] \cdots \Pr[H_0|\ln|\mathbf{X}(t-N+1)|, \mathbf{v}(t-N+1)] \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{2}. \quad (4.22)$$

Faisons un ordre de grandeur. En supposant que toutes ces probabilités *a priori* sont égales à p, ce test devient

$$p \underset{H_1}{\overset{H_0}{\gtrless}} \left(\frac{1}{2}\right)^{\frac{1}{N}}$$

ce qui s'interprète comme suit : pour que (4.22) soit vérifiée, chaque probabilité *a priori* doit individuellement (en ordre de grandeur) être supérieure à $(1/2)^{1/N}$, ce qui est un critère très contraignant dès que N augmente comme montré à la figure 4.1. Pour rendre ce test plus souple, on propose de remplacer (4.22) par

$$\Pr[H_0|\ln|\mathbf{X}(t)|, \mathbf{v}(t)] \cdots \Pr[H_0|\ln|\mathbf{X}(t-N+1)|, \mathbf{v}(t-N+1)] \underset{H_1}{\overset{H_0}{\gtrless}} q^N \quad (4.23)$$

où q est un seuil choisi arbitrairement, par exemple

$$q = \left(\frac{1}{2(N+1)}\right)^{\frac{1}{N}}. \quad (4.24)$$

Pour obtenir cette valeur, nous faisons un ordre de grandeur en supposant d'une part que nous décidons que la trame t est du silence si parmi les N trames $\{t, \cdots, t-$

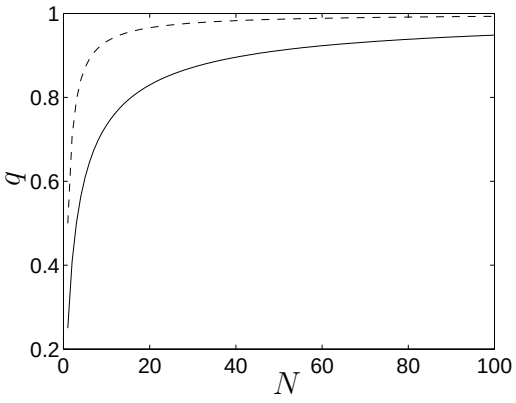


FIG. 4.1 – Influence de l'intégration : q (4.24) en fonction de N, le trait discontinu correspond à $(1/2)^{1/N}$.

$N+1\}$ au plus une est détectée comme étant de la parole, c'est-à-dire que

$$\begin{aligned}&\Pr[H_0|\ln\mathbf{X}(t),\mathbf{v}(t)]\cdots\Pr[H_0|\ln\mathbf{X}(t-N+1),\mathbf{v}(t-N+1)]\\&+\sum_{i=0}^{N-1}\Pr[H_1|\ln\mathbf{X}(t-i),\mathbf{v}(t-i)]\prod_{\substack{0\leq j\leq N-1\\ j\neq i}}\Pr[H_0|\ln\mathbf{X}(t-j),\mathbf{v}(t-j)]\underset{H_1}{\overset{H_0}{\gtrless}}\frac{1}{2}\end{aligned}$$

et d'autre part que toutes ces probabilités sont égales à q. Intégrer notre équation de décision sur les 10 dernières trames avec notre critère souple (4.23) nécessite en quelque sorte que chaque probabilité *a posteriori* $\Pr[H_0|\ln|\mathbf{X}(t-i)|,\mathbf{v}(t-i)]$ soit individuellement supérieure à 0.7341 contre 0.933 pour le critère (4.22).

Finalement, nous récapitulons le principe de notre détecteur audiovisuel d'activité vocale dans l'algorithme 1.

4.3 Détecteur visuel de silence

Dans le paragraphe précédent, nous avons proposé un détecteur audiovisuel d'activité vocale qui nécessite des connaissances statistiques sur le bruit (*i.e.* son allure spectrale donnée par Σ_B). Pour permettre une adaptation au cours du temps de celles-ci, une mise à jour de Σ_B est mise en œuvre (4.18). Cependant, celle-ci souffre d'un inconvénient majeur si le bruit est hautement non stationnaire puisque la rapidité d'adaptation de Σ_B dépend du paramètre d'adaptation γ. Pour surmonter cette difficulté, nous proposons une alternative qui consiste à exploiter uniquement la modalité visuelle de façon à introduire un détecteur *visuel* d'activité vocale. Par conséquent, ce détecteur est robuste à tout type d'environnement acoustique que se soit d'autres locuteurs, du bruit de fond non-stationnaire, *etc.* Plus particulièrement, nous allons chercher à distinguer les phases de silence pendant lesquelles le locuteur ne produit aucun son des autres phases actives de la parole.

Algorithme 1 Principe du détecteur audiovisuel d'activité vocale.

Pour tous les indices temporels t des trames **faire**
Calculer la TFCT $\mathbf{X}(t)$

/Calcul de la probabilité a posteriori *de* H_1*/*
Pour tous les noyaux $1 \leq i \leq N_{AV}$ **faire**
Estimer les facteurs d'amplitude $\hat{\alpha}_i(t)$ par l'algorithme itératif (4.15a)
Calculer les vraisemblances $p_G\big(\mathbf{v}(t)\,\big|\mu_i^V, \Sigma_i^V\big)\, p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\hat{\alpha}_i(t)\,\Gamma_i^A + \frac{\Sigma_B}{2}\big)$
Fin boucle
Calculer $\Pr[H_1|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$ par (4.16)

/Calcul de la probabilité a posteriori *de* H_0*/*
Calculer $\Pr[H_0|\mathbf{v}(t), \ln|\mathbf{X}(t)|]$ par (4.17)

/Equation de décision/
Intégrer l'équation de décision par (4.21) ou (4.23)

/Mise à jour des paramètres du silence/
Variance du bruit Σ_B par (4.18)
Paramètres vidéo du silence $\{\omega_i^{H_0}, \mu_i^{H_0}, \Sigma_i^{H_0}\}$ par (4.19a), (4.19b) et (4.19c)
Fin boucle

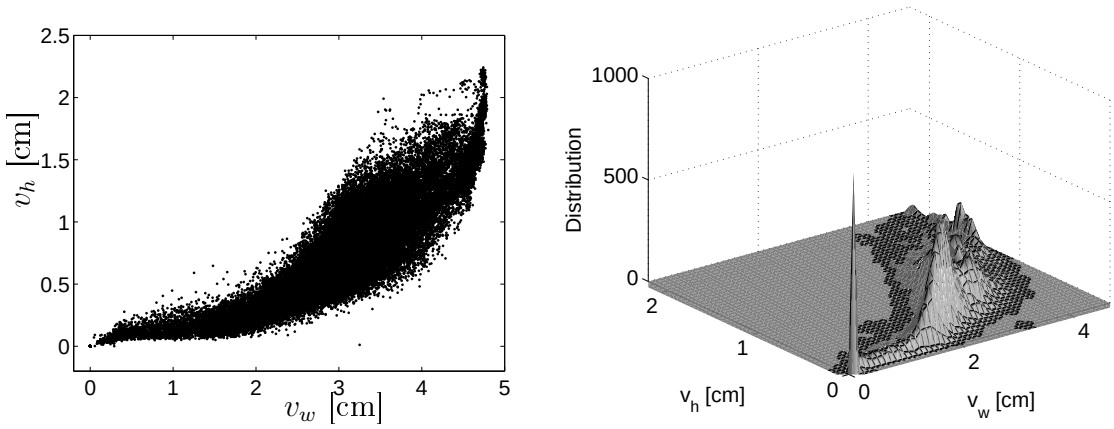
4.3.1 Principe de la détection visuelle d'activité vocale

L'idée centrale du détecteur visuel d'activité vocale est qu'en général, pendant la production de son, les lèvres bougent tandis qu'elles ne bougent pas (ou en tout cas beaucoup moins) pendant les silences. Il est nécessaire de recourir à des hypothèses dynamiques car il n'est pas possible d'exploiter directement la forme des lèvres de façon statique [119] : il est impossible à partir d'une seule image de déterminer si une personne parle ou ne parle pas. Ceci est confirmé par la distribution des paramètres visuels regroupant la hauteur et la largeur interne des lèvres (*cf.* figure 4.2). Ainsi, on peut voir qu'il n'y a pas de partition triviale entre les deux classes ("silence" et "non-silence") : en particulier, les lèvres fermées ne correspondent pas nécessairement à des moments de silence car elles sont présentes à la fois dans le silence et le non-silence. Il n'y a donc pas de relation directe entre des lèvres fermées et le silence ou bien entre une bouche ouverte et la parole : des considérations statiques sont bien insuffisantes pour caractériser le silence ou le non-silence.

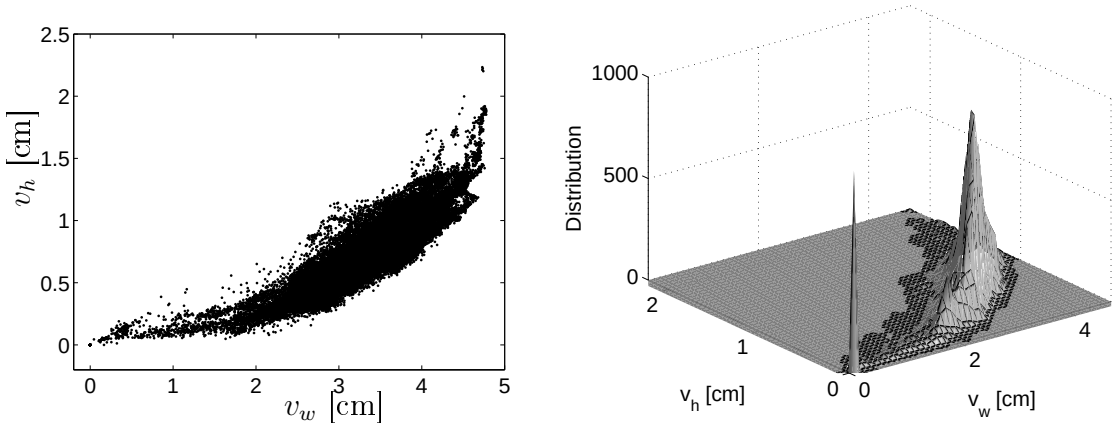
Ainsi, soit H_0 l'hypothèse "le locuteur ne produit aucun son" et H_1 "le locuteur produit un son". Nous proposons donc d'utiliser comme paramètre vidéo dynamique [119]

$$\pi(t) = \left|\frac{\partial v_w(t)}{\partial t}\right| + \left|\frac{\partial v_h(t)}{\partial t}\right| \tag{4.25}$$

où $v_w(t)$ et $v_h(t)$ sont respectivement les largeur et hauteur internes du contour labial. La classification entre le silence (H_0) et le non-silence (H_1) est fondée sur un seuillage : la trame à l'instant t est indexée comme du *silence* si $\pi(t)$ est inférieur à



(a) Non-silence



(b) Silence

FIG. 4.2 – Distribution des 129 500 paramètres vidéo dans le plan (hauteur,largeur) interne du contour labial pendant le silence et le non-silence. Sur les tracés de droite, les distributions ont été tronquées à 1000. Notons que 10% et 25% des points sont à l'origine (*i.e.* lèvres fermées) pour respectivement le non-silence et le silence.

un seuil λ et elle est indexée comme *non-silence* sinon

$$\pi(t) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{4.26}$$

Cependant, un seuillage direct de $\pi(t)$ ne s'avère pas très performant : par exemple, les lèvres peuvent être immobiles pendant plusieurs trames, alors que le locuteur est en train de parler (figure 4.3) : c'est notamment le cas des voyelles tenues (aux alentours de 1 seconde) ou les instants précédents l'ouverture rapide des lèvres pour le son [b] (aux alentours de 3 secondes). C'est pourquoi, $\pi(t)$ est d'abord lissé par intégration temporelle sur T trames consécutives :

$$\Pi(t) = \sum_{l=0}^{T-1} \alpha_l \, \pi(t-l) \tag{4.27}$$

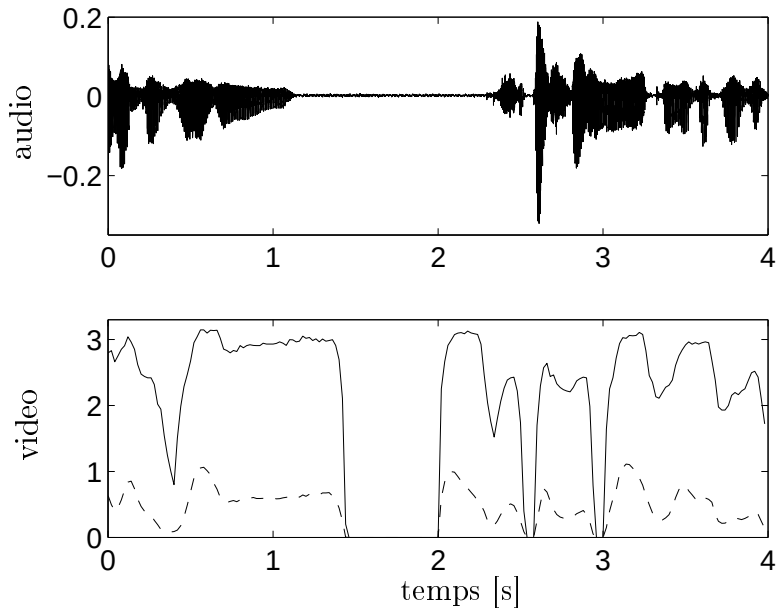


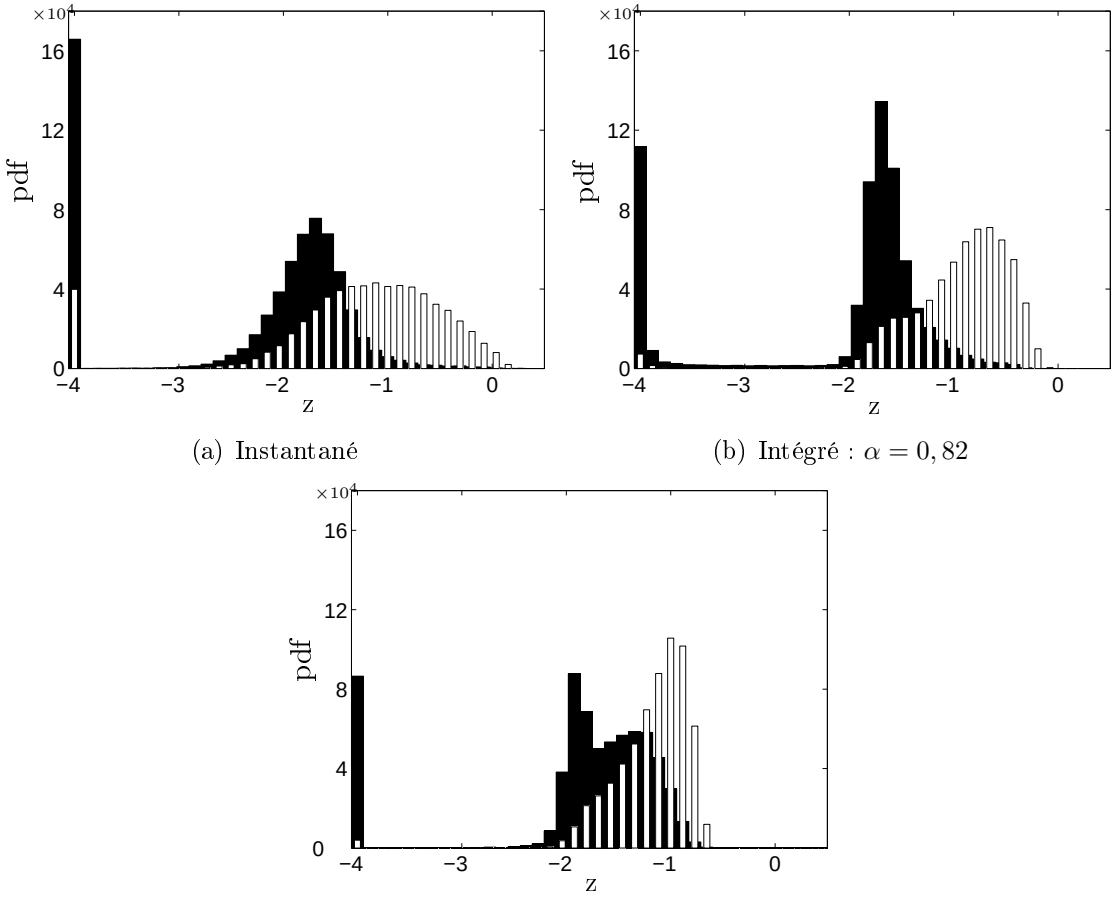
FIG. 4.3 – Détection visuelle de silence. La courbe du haut représente le signal acoustique. Les deux courbes du bas sont la largeur interne (trait continu) et la hauteur interne (trait discontinu) du contour labial.

où les α_l sont les coefficients d'un filtre passe-bas de réponse impulsionnelle infinie du premier ordre ($\alpha_l = \alpha^l$). La trame à l'instant t est alors classifiée comme *silence* si $\Pi(t)$ est inférieure à un nouveau seuil Λ et elle est classifiée comme *non-silence* sinon

$$\Pi(t) \underset{H_0}{\overset{H_1}{\gtrless}} \Lambda. \tag{4.28}$$

La figure 4.4 montre que le choix des paramètres d'intégration α_l doit être considéré avec attention. En effet, un choix d'un α trop petit (voir nul) conduit à une détection qui est sensible aux perturbations locales (petits mouvements des lèvres pendant les silences ou formes des lèvres stables pendant la parole) : les deux classes *silence* et *non-silence* sont alors largement superposées (*cf.* figure 4.4(a)) conduisant à un fort taux de fausses alarmes de détection de silence (*i.e.* décider *silence* alors que la trame correspondante est en réalité *non-silence*). Au contraire, choisir un α trop grand conduit à intégrer sur une trop longue durée : la fenêtre d'intégration englobe à la fois des trames de *silence* et de *non-silence*, conduisant ainsi à ne plus faire de distinction entre les deux classes (*cf.* figure 4.4(c)). Choisir un coefficient d'intégration acceptable permet de simplifier la classification comme le montre la figure 4.4(b).

Cependant, malgré un bon réglage de α, les deux classes *silence* et *non-silence* ne peuvent être totalement séparées. Il est impossible de trouver un seuil Λ parfait qui conduirait à détecter toutes les trames de silence sans produire de fausses alarmes de détection de silence. Dans notre problème de séparation de sources, on verra que



(a) Instantané

(b) Intégré : $\alpha = 0,82$

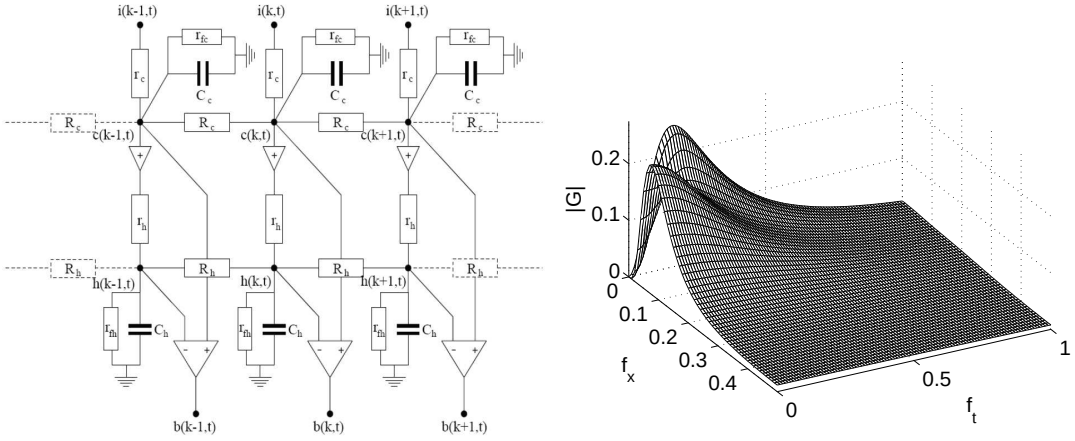
(c) Intégré : $\alpha = 0,99$

FIG. 4.4 – Influence de l'intégration sur le détecteur visuel d'activité vocale. Histogrammes du paramètre vidéo dynamique (4.27), en échelle logarithmique tronquée à -4 , pour trois valeurs des coefficients d'intégration $\alpha_l = \alpha^l$: instantané (a), valeur correcte de $\alpha = 0.82$ trames (b) et valeur trop grande de $\alpha = 0.99$ (c). Les histogrammes en noirs et blancs correspondent à $\Pi(t)$ défini par (4.27) pendant respectivement le silence et le non-silence.

l'important est de bien détecter les silences (*i.e.* avoir un taux de fausses alarmes relativement faibles). Dans ce cas, et comme pour le détecteur audiovisuel d'activité vocale, nous proposons de ne retenir comme périodes de silence que les sections composées d'au moins N trames consécutives individuellement détectées comme *silence*.

4.3.2 Détecteur visuel d'activité vocale sur images naturelles

Le détecteur visuel de silence que nous venons d'introduire fonctionne à partir des paramètres vidéo de largeur et hauteur internes du contour labial qui sont extraits par le système développé à l'ICP [80] que nous avons déjà évoqué au paragraphe 3.1. Bien qu'efficace, ce système n'en demeure pas moins lourd à mettre en œuvre : les lèvres doivent être maquillées en bleu (*cf.* figure 3.2 page 52) de façon à permettre



(a) (b)

FIG. 4.5 – Rétine artificielle. Figure 4.5(a) : schéma électrique équivalent de la rétine d'après [67, 11]. Figure 4.5(b) : réponse en fréquence spatio-temporelle de $G(z_s, f_t)$ (4.29) où $\alpha_c = 2$, $\beta_c = 0$, $\tau_c = 1$, $\alpha_h = 10$, $\beta_h = 0$, $\tau_h = 1$.

une segmentation plus facile des lèvres. Ainsi, pour s'affranchir de cet inconvénient et se rapprocher d'un système plus aisé à mettre en œuvre dans des conditions plus naturelles de prise de vue, nous allons maintenant introduire un nouveau détecteur de silence reposant également sur le principe d'un paramètrage vidéo dynamique mais exploitant directement des images brutes de la région des lèvres.

Réhaussement des contours labiaux

Nous proposons ainsi de faire une détection du contour labial fondée sur le fonctionnement de la rétine humaine [67, 11, 18] en exploitant un modèle électrique unidimensionnel de celle-ci. Un tel traitement permet entre autres de réhausser les contours, d'atténuer le bruit spatio-temporel et les variations de lumière.

Les photorécepteurs de la rétine humaine [88, 11] transforment l'intensité lumineuse de l'image perçue en un potentiel électrique $i(k,t)$ proportionnel à son logarithme (où k est l'indice du pixel considéré à l'instant t). On obtient ensuite des potentiels $b(k,t)$ en sortie des cellules OPL ("outer plexiform layer") de la rétine. Le schéma électrique équivalent modélisant le lien entre $i(k,t)$ et $b(k,t)$ est donné à la figure 4.5(a) [67, 11]. Il s'agit d'un filtre spatio-temporel non séparable dont la fonction de transfert dans l'espace de Fourier pour sa partie temporelle et dans l'espace de la transformée en Z pour sa partie spatiale est donnée par

$$G(z_s, f_t) = \frac{B(z_s, f_t)}{I(z_s, f_t)}$$

où $B(z_s, f_t) = TZ_s\{TF_t\{b(k,t)\}\}$ (resp. $I(z_s, f_t) = TZ_s\{TF_t\{i(k,t)\}\}$) est la transformée en Z spatiale (notée $TZ_s\{\cdot\}$) de la transformée de Fourier temporelle (notée

(a) Avant traitement (b) Après traitement

FIG. 4.6 – Illustration du traitement effectué par la rétine.

$TF_t\{\cdot\}$) de $b(k,t)$ (resp. $i(k,t)$). On peut montrer [11] que

$$G(z_s, f_t) = \frac{1}{1+\beta_c+\alpha_c\left(-z_s^{-1}+2-z_s\right)+\jmath 2\pi f_t\tau_c} \times \frac{\beta_h+\alpha_h\left(-z_s^{-1}+2-z_s\right)+\jmath 2\pi f_t\tau_h}{1+\beta_h+\alpha_h\left(-z_s^{-1}+2-z_s\right)+\jmath 2\pi f_t\tau_h} \quad (4.29)$$

avec $\alpha_c = r_c/R_c$, $\beta_c = r_c/r_{fc}$, $\tau_c = r_c C_c$, $\alpha_h = r_h/R_h$, $\beta_h = r_h/r_{fh}$, $\tau_h = r_h C_h$. Dans cette expression, α_h (resp. α_c) représente la constante d'espace des cellules $h(k,t)$ (resp. $c(k,t)$), β_h et β_c représentent leurs constantes de fuite et τ_h et τ_c leurs constantes de temps. La figure 4.5(b) montre sa réponse en fréquence spatio-temporelle. On constate que ce filtre présente un comportement passe-bande spatial pour les faibles fréquences temporelles qui tend à devenir passe-bas quand la fréquence temporelle augmente. De façon duale, ce filtre a un comportement passe-bande temporel pour les faibles fréquences spatiales qui tend à devenir passe-bas quand la fréquence spatiale augmente. Une illustration du traitement effectué par la rétine est présentée à la figure 4.6. Cette structure de filtre permet un implémentation rapide de la détection de contours [11, 18].

Détecteur visuel d'activité vocale sur images naturelles

Une fois le traitement rétinien de réhaussement des contours labiaux effectué, nous appliquons une transformée de Fourier bidimensionnelle à chaque image résultante $r(t) \in \mathbb{R}^{N_u \times N_v}$ (où N_u et N_v sont respectivement le nombre de lignes et colonnes de l'image $r(t)$) :

$$R_{uv}(t) = \sum_{l=0}^{N_u-1}\sum_{c=0}^{N_v-1} r_{lc}(t)\, w_{lc}\, e^{-\jmath 2\pi\left(l\frac{u}{N_l}+c\frac{v}{N_c}\right)}. \quad (4.30)$$

où w_{lc} est la fenêtre de Hamming bidimensionnelle. Par la suite, on ne garde que le carré du module $|R(t)|^2$, où $R(t)$ est la matrice rassemblant les termes $R_{uv}(t)$, afin

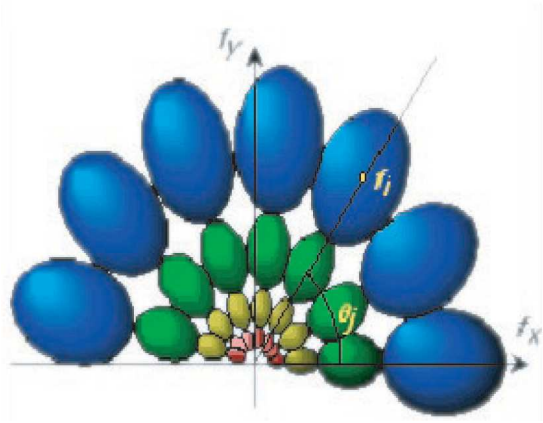


FIG. 4.7 – Transformation log-polaire d'après [19], chacun des ovales correspond au filtre de Gabor log-polaire centré sur la fréquence spatiale f_i dans la direction θ_j.

de détecter les mouvements des lèvres au cours du temps. Pour cela, nous proposons d'effectuer une opération de dérivation temporelle :

$$\Delta R(t) = \left| |R(t)|^2 - |R(t-1)|^2 \right|. \tag{4.31}$$

Le fait de recourir ici au module de la transformée de Fourier bidimensionnelle permet de s'affranchir largement des mouvements parasites de translations du visage par rapport à la caméra qui n'ont une influence notable que dans la phase. Cette opération de dérivation temporelle est suivie d'un filtrage de type passe-bande spatial de façon à atténuer les effets du bruit et des variations de lumière. Pour cela, nous effectuons une transformation log-polaire (*cf.* figure 4.7) de $\Delta R(t)$ et nous ne gardons, pour toutes les directions θ_i, que certaines fréquences spatiales f_k ce qui nous donne $\Delta R^F(t)$. Cette transformation log-polaire est calculée à l'aide de filtres de Gabor log-polaires $G_{ik}(f,\theta)$ centrés à la fréquence f_k dans la direction θ_i [64] :

$$G_{ik}(f,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \left(\frac{f_k}{f}\right)^2 \exp\left(-\frac{\ln\left(\frac{f}{f_k}\right)^2}{2\sigma^2}\right) \cos\left(\frac{1+\cos(\theta-\theta_i)}{2}\right)^{50}$$

où σ est un facteur d'échelle.

Finalement, le paramètre vidéo dynamique $\pi(t)$ que nous allons utiliser pour notre détecteur visuel d'activité vocale sur images naturelles est défini comme la moyenne sur les lignes et les colonnes de $\Delta R^F(t)$:

$$\pi(t) = \frac{1}{N_u}\frac{1}{N_v} \sum_{u=0}^{N_u-1} \sum_{v=0}^{N_v-1} \Delta R^F_{uv}(t). \tag{4.32}$$

Le détecteur visuel de silence sur images naturelles classe alors comme *silence* la trame à l'instant t si $\pi(t)$ est inférieur à un seuil λ et comme *non-silence* sinon :

$$\pi(t) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{4.33}$$

Il est également possible de recourir à une intégration temporelle de façon à améliorer les performances du détecteur

$$\Pi(t) = \sum_{n=0}^{T-1} \alpha_l \, \pi(t-n) \underset{H_0}{\overset{H_1}{\gtrless}} \Lambda \tag{4.34}$$

où comme précédemment les α_l sont les coefficients d'un filtre passe-bas de réponse impulsionnelle infinie du premier ordre ($\alpha_l = \alpha^l$).

Nous résumons le principe global du détecteur visuel d'activité vocale sur images naturelles dans l'algorithme 2.

Algorithme 2 Détecteur visuel d'activité vocale sur image naturelle.

Pour tous les indices temporels t **faire**

/Réhaussement des contours par filtrage rétinien/
Calcul de $r(t)$ obtenue par le filtrage (4.29)

/Calcul du paramètre vidéo dynamique/
Calcul de la transformée de Fourier bi-dimensionelle de $r(t)$ (4.30)
Dérivation temporelle de $R(t)$ par (4.31)
Filtrage spatial par transformation log-polaire pour obtenir $\Delta R^F(t)$
Calcul du paramètre dynamique instantané $\pi(t)$ par (4.32)

/Equation de décision/
Intégrer l'équation de décision par (4.34)

Fin boucle

4.4 Corpus

Les deux corpus présentés au chapitre 3 présentent l'intérêt d'être bien contrôlés et de bien représenter la richesse audiovisuelle de la parole. Cependant, ils ne contiennent pas de périodes naturelles de silence pendant lesquelles les locuteurs ne parlent pas. Pour cette raison, nous considérons dans cette partie deux nouveaux corpus comportant différents locuteurs en interaction. Nous avons utilisé d'une part un corpus enregistré à l'ICP au début de notre étude avec David Sodoyer et Jean-Luc Schwartz (que nous appellerons "corpus Grenoble") et d'autre part un corpus que nous avons enregistré avec Andrew Aubrey et Yulia Hicks à Cardiff dans le laboratoire "Center of Digital Signal Processing" dirigé par le professeur Jonathon Chambers de l'université de Cardiff aux Pays de Galles (que nous appellerons "corpus Cardiff").

4.4.1 Corpus "Grenoble"

Ce corpus a été enregistré à l'ICP avec des moyens audiovisuels communs au LIS et à l'ICP. L'enregistrement a été supervisé par David Sodoyer et Christophe Savariaux. La figure 3.2 page 52 montre le montage utilisé pour cet enregistrement :

Nom	Langue maternelle
loc. 1	Cantonnais
loc. 2	Cantonnais
loc. 3	Mandarin
loc. 4	Mandarin
loc. 5	Indien
loc. 6	Persan
loc. 7	Grec
loc. 8	Arabe
loc. 9	Créole
loc. 10	Russe
loc. 11	Français

TAB. 4.1 – Liste des locuteurs avec leur langue maternelle.

les deux locuteurs ont chacun une micro caméra fixée sur un casque et focalisée sur la région des lèvres. Celles-ci sont maquillées en bleue de façon à permettre leur segmentation par le système développé à l'ICP [80] présenté précédemment. Ce corpus sera donc utilisé pour tester les techniques du paragraphe 4.2 portant sur les paramètres hauteur et largeur du contour labial. Les deux sujets étaient placés dans des pièces séparées de façon à pouvoir enregistrer chacun d'eux dans des conditions contrôlées permettant d'obtenir des signaux acoustiques propres (*i.e.* où seul un locuteur est présent à la fois).

Les deux locuteurs, dont le français est la langue maternelle, ont été placés dans diverses situations de dialogue spontané : devinettes, dialogues sur des sujets fournis, jeux interactifs, *etc.* Ces diverses situations regroupent des silences plus ou moins longs (hésitations, réflexions, *etc*) des accélérations dans la conversation, des coupures de parole, *etc.* Ce corpus représente un total d'environ 43 minutes de parole spontanée soit environ 129 000 trames audiovisuelles comprenant environ 50% de trames de silence. L'indexation manuelle des trames, entre trames de silence et les autres, a été réalisée par David Sodoyer.

4.4.2 Corpus "Cardiff"

Un deuxième corpus comprenant également des silences dans de la parole spontanée a été enregistré pour notre étude lors d'un échange PAI Alliance avec le laboratoire "Center of Digital Signal Processing" de l'université de Cardiff aux Pays de Galles. Ce corpus est destiné à être utilisé pour la détection d'activité vocale sur images naturelles du paragraphe 4.3. Il fait intervenir 11 sujets de langue maternelle différente (*cf.* tableau 4.1). Lors de chaque enregistrement, les locuteurs étaient placés seuls dans une pièce comprenant des caméras fixes enregistrant le visage de face ainsi qu'une vue de côté (*cf.* figure 4.8). Dans le cadre du projet TELMA (Terminal de téléphonie à l'usage des malentendants), la vue de côté doit permettre de tester la possibilité de faire de la détection d'activité vocale en simulant ce que pourrait enregistrer une caméra fixée à une oreillette filmant la zone des lèvres du locuteur

(a) Vue de face (b) Vue de côté

FIG. 4.8 – Exemple d'enregistrement pour trois locuteurs avec la vue de face et celle de côté.

de côté. Les deux caméras sont des caméras "firewire" à 30 images par seconde, synchrones entre elles et de résolution 480×680 pixels. Les locuteurs étaient assis sur une chaise devant une table sur laquelle se trouvait un écran d'ordinateur leur donnant les instructions à suivre ainsi que les tâches à effectuer. De façon à limiter les mouvements de la tête les sujets avaient pour instruction de l'appuyer contre le mur derrière eux (cette consigne n'ayant pas toujours été respectée tout au long des enregistrements, ceux-ci ne sont pas exploitables dans leur totalité).

Pour chaque sujet deux enregistrements ont été faits : le premier en langue anglaise et le second dans leur langue maternelle. Chaque enregistrement comporte deux types de tâches dont les instructions correspondantes sont écrites en anglais même si le sujet doit répondre dans sa langue maternelle. La première tâche consiste à répondre en quelques phrases à des questions banales telles que "Quel est le dernier livre que vous avez lu ?" ou à effectuer des opérations de calcul mental comme par exemple "$\left(\frac{111}{3} - 7\right) \times 8$". Dans la seconde tâche, le sujet voit inscrit sur l'écran un nom de couleur (par exemple vert) écrit dans une couleur pouvant être différente (par exemple bleue) : "VERT". Le sujet doit alors dire "Le mot est vert, la couleur est bleue". Le but recherché par ces tests est simplement que le sujet ne pense plus au fait qu'il est enregistré de façon à ce que les attitudes et la parole soient spontanées et non pas contrôlées.

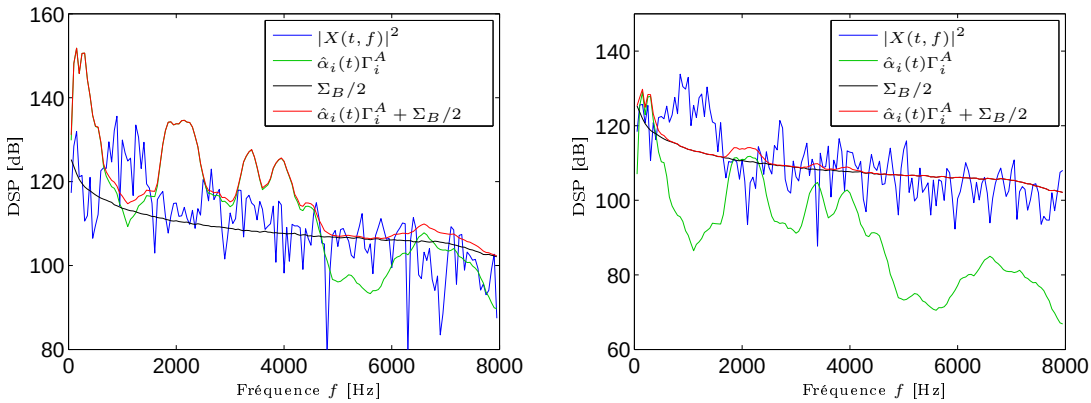
Les différences entre ce corpus et le corpus enregistré à Grenoble sont principalement d'une part le fait qu'ici les lèvres des sujets ne sont pas maquillées en bleu et d'autre part que la position relative de la zone des lèvres et les caméras ne sont pas fixes. Cette dernière différence est d'ailleurs à l'origine de plusieurs problèmes : les sujets n'ayant pas nécessairement gardé la tête fixe, les enregistrements comportent des mouvements de tête nuisibles à la détection de l'activité vocale.

4.5 Expérimentations

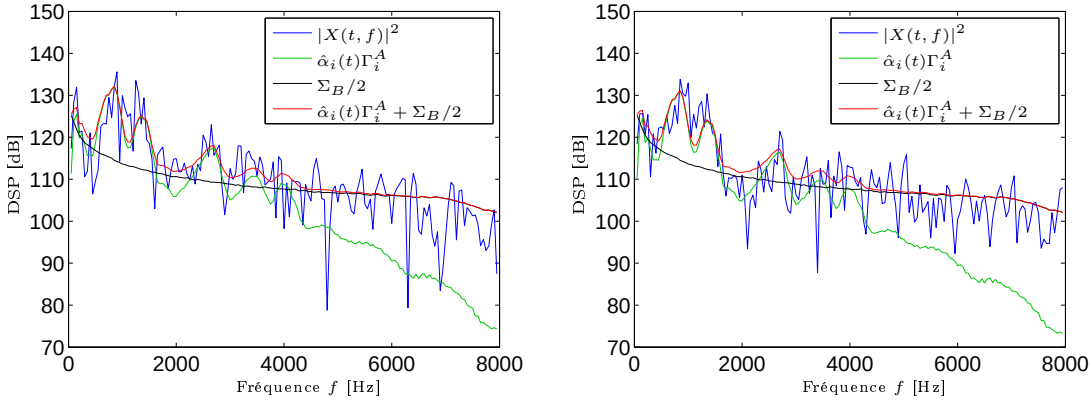
Dans ce paragraphe, nous présentons tout d'abord les résultats expérimentaux de la détection d'activité vocale par le modèle audiovisuel, puis les résultats expérimentaux de la détection de silence purement visuelle.

4.5.1 Détecteur audiovisuel d'activité vocale

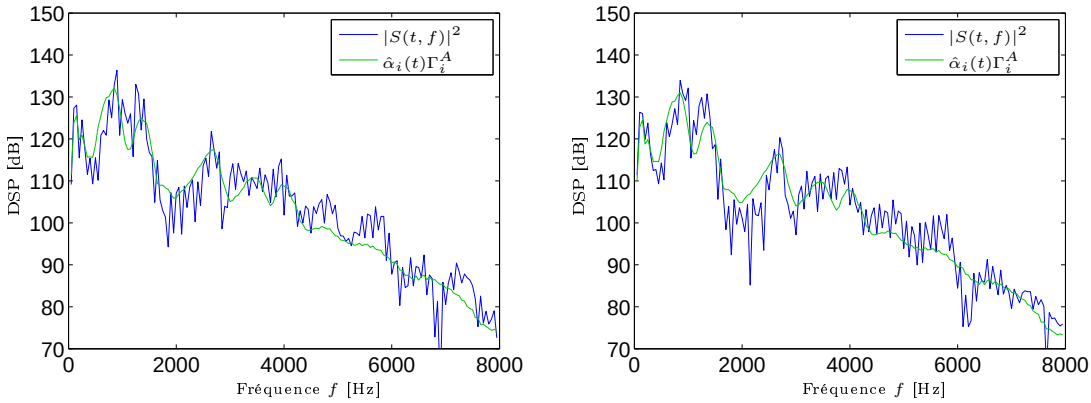
Dans une première expérience, nous présentons les résultats concernant l'estimation du facteur d'amplitude (paragraphe 4.2.2). Nous considérons ici le corpus des logatomes associé au modèle audiovisuel à 12 noyaux dont les paramètres ont été appris par l'algorithme EM (*cf.* figure 3.9). Pour tester l'estimation du facteur d'amplitude $\hat{\alpha}(t)$ par l'algorithme itératif (4.15a), nous sélectionnons parmi le corpus des logatomes, le logatome [a] qui défini ainsi le signal $s(t) = \sqrt{\alpha(t)}s'(t)$. Pour créer l'observation bruitée $x(t)$ (4.10), nous lui ajoutons un bruit coloré $b(t)$ dont nous déterminons la matrice de covariance spectrale Σ_B. Nous choisissons ensuite deux matrices de localisations Γ_i^A, caractéristiques de l'allure spectrale du son modélisé, correspondant au son [a] (noyau numéroté 12 de notre modèle) et au son [y] (noyau numéroté 4 de notre modèle). Finalement, pour ces deux matrices de localisation, nous estimons le facteur d'amplitude $\hat{\alpha}(t)$. Les résultats obtenus sont présentés à



(a) Noyau /y/ (b) Noyau /y/



(c) Noyau /a/ (d) Noyau /a/



(e) Noyau /a/ (f) Noyau /a/

FIG. 4.9 – Estimation du facteur d'amplitude pour deux noyaux différents /a/ et /y/ et pour deux trames du signal $s(t)$ (colonne de gauche et de droite). Figures 4.9(a), 4.9(b) (resp. 4.9(c), 4.9(d)) estimation de $\alpha_i(t)$ pour le noyau /y/ (resp. /a/). Figures 4.9(e) et 4.9(f) spectre du signal non bruité $s(t)$ et son estimation à partir du modèle $\hat{\alpha}_i(t)\Gamma_i^A$.

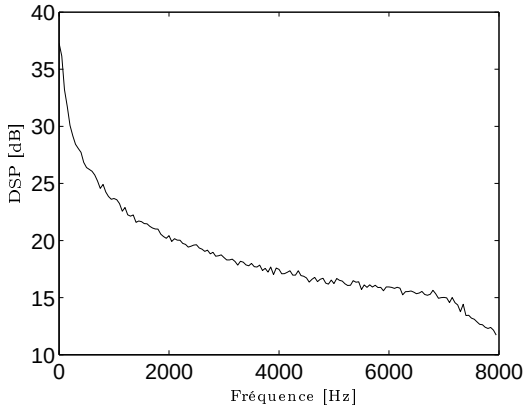


FIG. 4.10 – Densité spectrale de puissance du bruit.

la figure 4.9. La colonne de gauche (figures 4.9(a), 4.9(c), 4.9(e)) correspond à une première trame du signal $s(t)$ et celle de droite (figure 4.9(b), 4.9(d), 4.9(f)) à une seconde trame du signal $s(t)$. Les figures 4.9(c) et 4.9(d) représentent, pour les deux trames, les résultats de l'estimation du facteur d'amplitude $\alpha_i(t)$ associé à la matrice de localisation Γ_i^A correspondant au noyau modélisant le [a]. Comme on peut le constater, l'estimation de $\hat{\alpha}_i(t)$ permet de bien modéliser l'allure spectrale du son bruité comme le montre le tracé de $\hat{\alpha}_i(t)\Gamma_i^A + \Sigma_B/2$ (courbe rouge). La vraisemblance $p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\alpha(t)\,\Gamma_i^A + \Sigma_B/2\big)$ a pour ordre de grandeur 10^{-68}, ce qui correspond en moyenne à une vraisemblance marginale $p_{LR}\big(\ln|X(t,f)|\,\big|\alpha(t)\,\Gamma_i^A(f) + \Sigma_B(f)/2\big)$, à chaque fréquence f, de l'ordre de 0.38 : les coefficients $\ln|X(t,f)|$ sont donc proches du mode de la loi LogRayleigh (*cf.* figure 3.4 page 57). En revanche, dans le cas de la matrice de localisation Γ_i^A associée au noyau modélisant le son [y], bien que $\hat{\alpha}_i(t)$ maximise la vraisemblance $p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\alpha(t)\,\Gamma_i^A + \Sigma_B/2\big)$, la matrice spectrale $\hat{\alpha}_i(t)\Gamma_i^A + \Sigma_B/2$ ne permet pas de modéliser correctement le signal bruité comme le montre les figures 4.9(a) et 4.9(b). Dans ce cas, la vraisemblance $p_{LR}\big(\ln|\mathbf{X}(t)|\,\big|\alpha(t)\,\Gamma_i^A + \Sigma_B/2\big)$ a une valeur dont l'ordre de grandeur est 10^{-230}, ce qui correspond en moyenne à une vraisemblance marginale à chaque fréquence de l'ordre de 0.035 : les coefficients $\ln|X(t,f)|$ sont alors plus éloignés du mode de la loi LogRayleigh. Finalement, nous avons représenté aux figures 4.9(e) et 4.9(f) les coefficients de la TFCT du signal non bruité $S(t,f)$ (courbe bleue) ainsi que la matrice spectrale $\hat{\alpha}_i(t)\Gamma_i^A$, où Γ_i^A est la matrice de localisation du noyau modélisant le son [a]. La matrice $\hat{\alpha}_i(t)\Gamma_i^A$ correspond bien à une estimation de l'enveloppe spectrale du son non bruité $S(t,f)$, ce qui confirme la bonne estimation du facteur d'amplitude $\hat{\alpha}_i(t)$.

Pour tester ensuite notre détecteur d'activité vocale audiovisuel (algorithme 1), le signal $s(t)$, dont nous cherchons à détecter la présence ou non, est issu du corpus "Grenoble", le bruit est un bruit coloré dont la densité spectrale de puissance est représentée à la figure 4.10. Nous avons utilisé 10% des trames indexées manuellement *parole* pour l'apprentissage du modèle audiovisuel (3.18) dont l'ensemble des paramètres $\{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Gamma_i^A\}_{1\leq i\leq N_{AV}}$ ont été calculés en utilisant les résultats du paragraphe 3.3.4. Dans nos expériences, nous avons choisi un nombre de noyaux

N_{AV} égal à 65 réalisant ainsi un compromis entre bonne modélisation et complexité du modèle. Les performances de notre détecteur d'activité vocale sont représentées par les courbes COR (caractéristiques opérationnelles de réception). Elles correspondent au taux de bonnes détection des silences en fonction du taux de fausses alarmes. Le taux de bonne détection est défini comme le rapport entre le nombre de trames détectées *silence* alors qu'elles correspondent effectivement à du silence $N_{H_0|H_0}$ et le nombre de trames indexées manuellement *silence* pris pour référence N_{H_0}

$$BD = \frac{N_{H_0|H_0}}{N_{H_0}}. \tag{4.35}$$

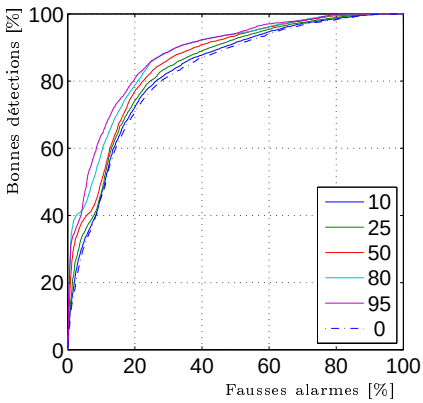
Le taux de fausses alarmes, défini à partir du nombre de trames détectées *non silence* correspondant effectivement à des trames de non silence $N_{H_1|H_1}$, est donné par

$$FA = \frac{N_{H_0|H_1}}{N_{H_1}} = 1 - \frac{N_{H_1|H_1}}{N_{H_1}}, \tag{4.36}$$

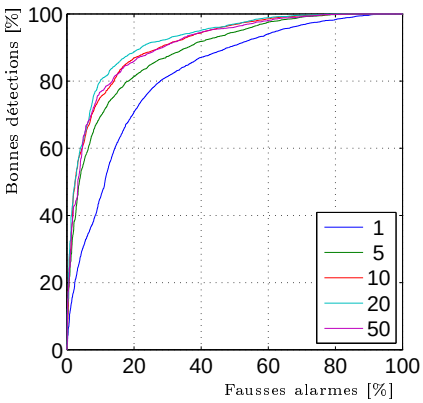
où N_{H_1} est le nombre de trames indexées *non silence* manuellement pris pour référence.

Tout d'abord, intéressons-nous à l'influence de l'intégration discutée au paragraphe 4.2.4 sur les performances pour différents rapport signal sur bruit (RSB). Les résultats sont présentés à la figure 4.11 où nous avons testé les deux intégrations (4.20b) et (4.23) décrites respectivement par le paramètre κ et le nombre de trames d'intégration N. D'une manière générale, nous constatons d'une part que l'intégration permet d'améliorer les performances du détecteur d'activité vocale comme le montrent les courbes COR et ce quelque soit le RSB et d'autre part, que plus le RSB est important meilleures sont les performances. Concernant l'intégration (4.20b), il est intéressant de noter qu'augmenter le paramètre κ permet, dans une certaine mesure, d'améliorer les performances : par exemple, à un RSB de 0dB, pour 80% de bonnes détections, le taux de fausses alarmes passe de 18%, dans le cas instantané, à 9% avec $\kappa = 0.8$. Cependant, choisir κ trop grand peut détériorer les performances : par exemple dans le cas d'un RSB de 10dB, les performances sont moins bonnes avec $\kappa = 0.95$ que pour $\kappa = 0.8$. Concernant l'intégration (4.23), les figures 4.11(b), 4.11(d) et 4.11(f) montrent l'importance du choix de la durée d'intégration N. En effet, choisir une durée de l'ordre de 20 trames permet de considérablement améliorer les performances quelque soit le RSB. En revanche, choisir une durée d'intégration trop longue ($N = 50$ par exemple) conduit à diminuer le taux de détection des silences. On constate d'une manière générale que les performances du détecteur d'activité vocale sont meilleures dans le cadre de l'intégration proposée (4.23) que pour l'intégration (4.20b). Ceci est d'autant plus significatif que le RSB est faible comme on peut le voir sur les figures 4.11(a) et 4.11(b).

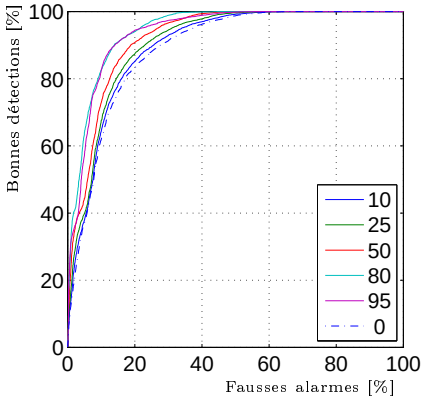
Dans une seconde série d'expériences (figure 4.12), nous comparons trois détecteurs d'activité vocale : le détecteur audiovisuel d'activité vocale fondé sur le modèle audiovisuel multi-noyaux (AV), le détecteur d'activité vocale purement audio fondé sur le modèle purement audio multi-noyaux obtenu par marginalisation du modèle audiovisuel (A), le détecteur d'activité vocale fondé sur un modèle purement audio global [121] (A Global). Dans ces trois cas, nous exploitons l'intégration (4.23) avec une durée d'intégration de 400ms, soit $N = 20$, déduite de l'étude précédente. Sur la



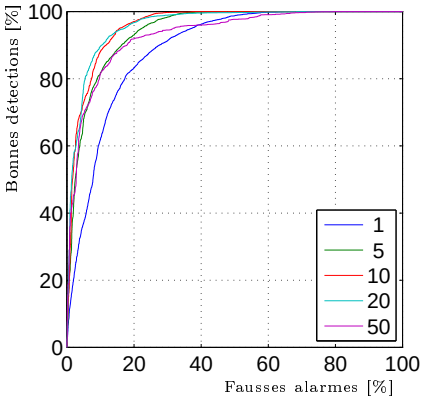
(a) RSB = −10dB



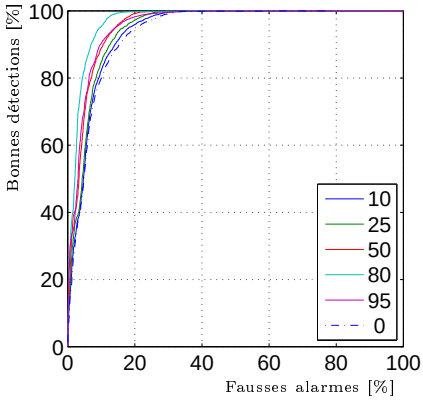
(b) RSB = −10dB



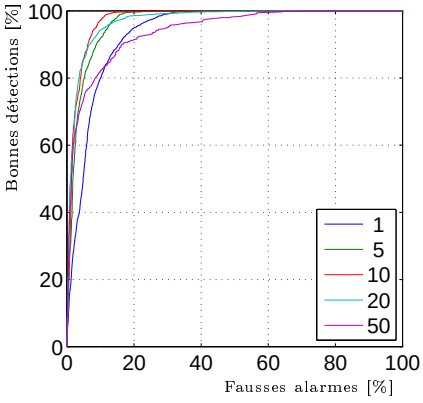
(c) RSB = 0dB



(d) RSB = 0dB

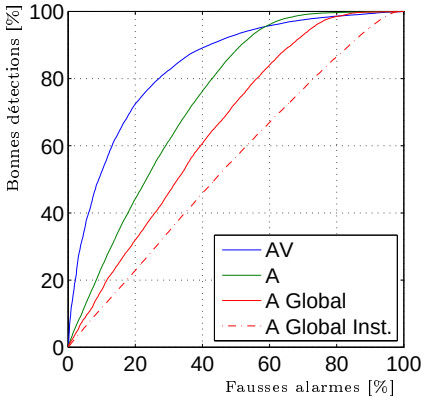


(e) RSB = 10dB

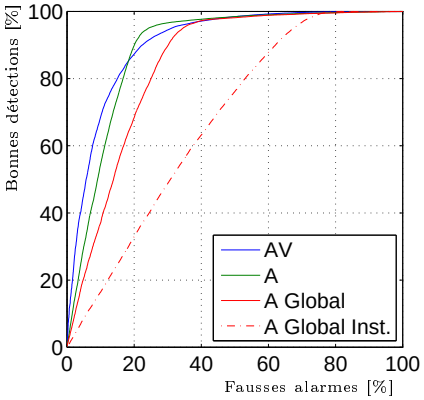


(f) RSB = 10dB

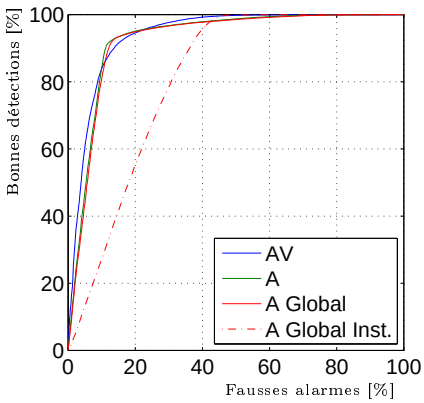
FIG. 4.11 – Influence de l'intégration (4.20b) (figures 4.11(a), 4.11(c) et 4.11(e)) ou (4.23) (figures 4.11(b), ou 4.11(d) et 4.11(f)) sur les courbes COR pour différents RSB. Figures 4.11(a), 4.11(c) et 4.11(e) : les légendes correspondent à 100κ (0 étant les probabilités instantanées sans intégration). Figures 4.11(b), ou 4.11(d) et 4.11(f) : les légendes correspondent aux nombres de trames d'intégration N.



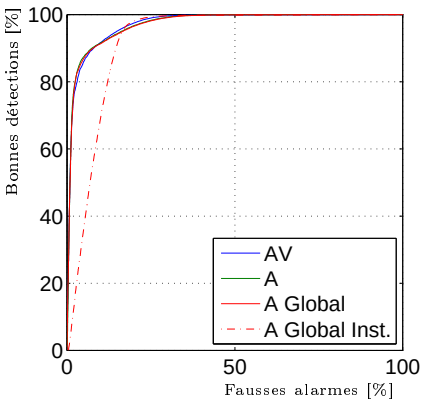
(a) RSB = -10 dB



(b) RSB = 0 dB

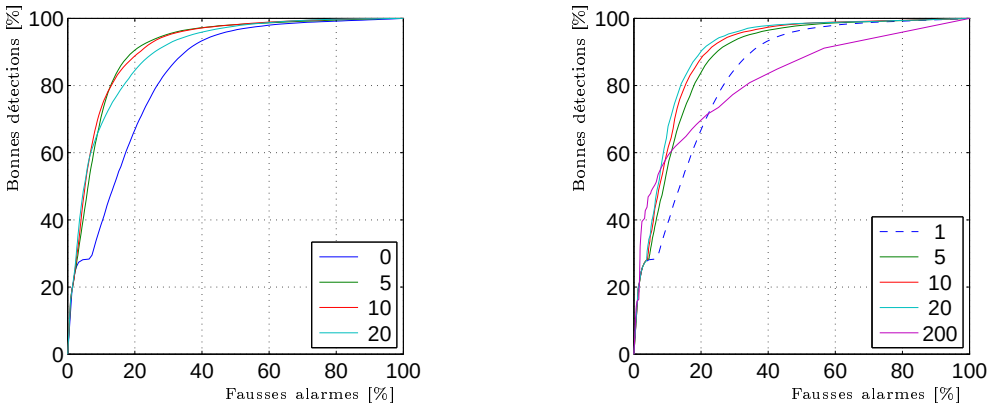


(c) RSB = 10 dB



(d) RSB = 30 dB

FIG. 4.12 – Comparaison des détecteurs d'activité vocale audiovisuel (AV), audio (A), audio global (A Global) et audio global instantané (A Global Inst.) pour différents RSB.



(a) Influence de l'intégration (b) Influence de la durée minimale des silences

FIG. 4.13 – Performances du détecteur visuel de silence à partir des paramètres de largeur et hauteur internes. Figure 4.13(a) : influence du filtrage passe-bas sur les performances (4.27), la légende indique la valeur de τ. Figure 4.13(b) : influence de la durée minimale des silences, la légende indique le nombre minimal de trames consécutives de silence que l'on peut détecter.

figure 4.12, nous avons également reporté les performances obtenues par le détecteur d'activité vocale [121] dans le cas instantané (A Global Inst.). D'une façon générale, il est possible d'établir la relation d'ordre suivante pour la détection des silences :

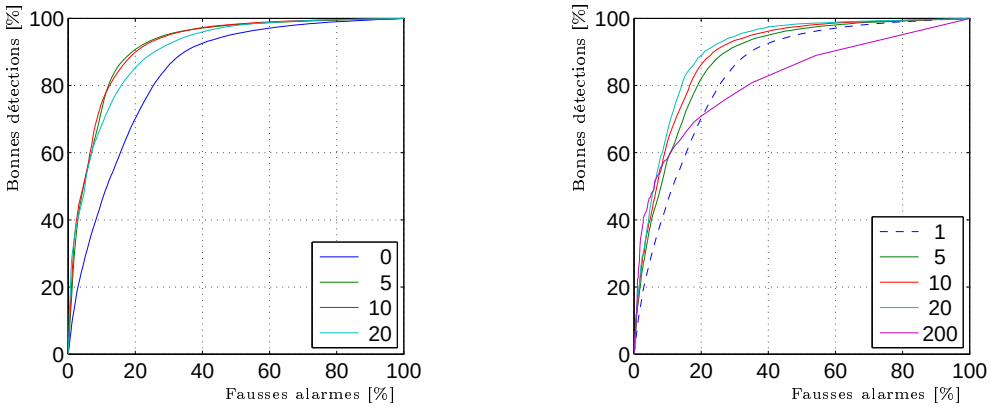
$$\text{(A Global Inst.)} < \text{(A Global)} < \text{(A)} < \text{(AV)}.$$

Ce classement est d'autant plus pertinent que le RSB est faible. En effet, alors que l'amélioration apportée par la modalité visuelle est quasi imperceptible pour un RSB de 10dB ou 30dB, elle permet de passer d'un taux de fausses alarmes de presque 60%, pour le détecteur (A Global), à environ 25%, pour le détecteur (AV), pour 80% de bonnes détections quand le RSB est de -10dB. Ainsi, plus le RSB est faible plus la redondance de la modalité visuelle de la parole permet de suppléer le manque d'information acoustique directement exploitable pour la détection d'activité vocale. De plus, les performances obtenues par le détecteur d'activité vocale (A) sont meilleures que celles obtenues par le modèle global (A Global), ce qui tend à prouver que le modèle muti-noyaux permet de mieux modéliser la parole. Finalement, ces figures permettent de montrer le gain combiné de la modalité visuelle et de l'intégration par rapport à un modèle purement acoustisque instantané, surtout pour un faible RSB.

4.5.2 Détecteur visuel de silence

Dans ce paragraphe, nous présentons les résultats du détecteur visuel de silence tout d'abord pour les paramètres de largeur et hauteur internes du contour labial, puis pour les images naturelles.

Considérons dans un premier temps le corpus "Grenoble" et donnons les performances du détecteur de silence à partir des paramètres de largeur et hauteur internes

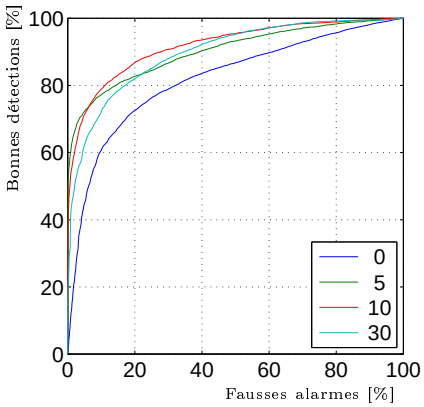


(a) Influence de l'intégration (b) Influence de la durée minimale des silences

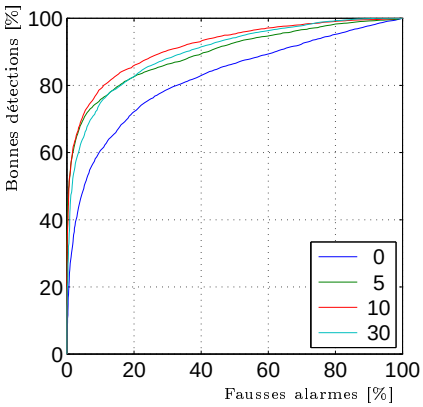
FIG. 4.14 – Performances du détecteur visuel de silence sur images naturelles pour le corpus "Grenoble". Figure 4.14(a) : influence de l'intégration (4.27). La légende indique la valeur de τ. Figure 4.14(b) : influence de la durée minimale des silences détectables. La légende indique la durée des silences en nombre de trames.

des lèvres (figure 4.13). Nous pouvons voir l'importance de l'intégration par le filtrage passe-pas (4.27) dans l'amélioration des performances du détecteur de silence (figure 4.13(a)) : cette intégration diminue l'influence des faibles mouvements des lèvres pendant la parole (FA) et diminue également l'influence des petits mouvements des lèvres pendant les silences (ND). Ainsi, cela permet de diminuer de façon significative le taux de fausses alarmes à un taux de bonnes détections donné par rapport à l'utilisation du paramètre instantané (4.25) : par exemple le point 28%-80% sans intégration devient 12%-80% avec une intégration correcte ($\tau = 5$). De plus, la figure 4.13(b) montre l'effet de ne détecter que des silences d'au moins une certaine durée égale à N trames consécutives de silence. Les courbes COR montrent que choisir une durée minimale de silence trop large ($N = 200$ trames ce qui correspond à 4s) diminue de façon importante le taux de détection des silences. En revanche, choisir de façon raisonnable une durée minimale des silences détectés (par exemple $N = 20$ trames ce qui correspond à 400ms) permet de diminuer le nombre de fausses détections tout en conservant un fort taux de détection. Le gain de performance dû à la durée minimale des silences est similaire à celui que l'on peut obtenir en intégrant par le filtrage passe-bas. Utilisé simultanément, nous obtenons des scores de détection de silence qui sont exploitables pour la séparation de sources comme nous le verrons dans la partie suivante. Le compromis 10%-70% peut garantir un taux de fausses alarmes suffisamment faible pour une exploitation correcte de la détection des silences.

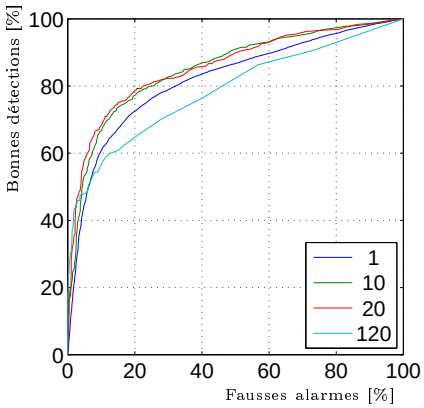
Considérons maintenant le cas du détecteur de silence sur images naturelles. Tout d'abord, nous avons utilisé le corpus "Grenoble" pour faire une comparaison entre les performances obtenues par le détecteur visuel de silence (4.27) (figure 4.13) et celles obtenues par le détecteur de silence sur images naturelles (figure 4.14). Comme pour le détecteur visuel de silence obtenu à partir des largeur et hauteur internes des lèvres, l'intégration par un filtre passe-bas du paramètre vidéo (4.32) ou la durée minimale des silences détectable permettent d'améliorer la robustesse de la détection



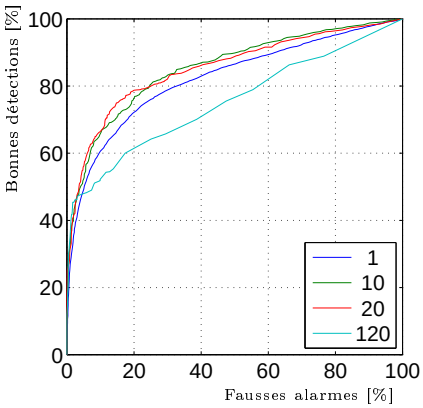
(a) Influence de l'intégration : vue de face



(b) Influence de l'intégration vue de côté



(c) Influence de la durée minimale des silences : vue de face



(d) Influence de la durée minimale des silences : vue de côté

FIG. 4.15 – Performances du détecteur visuel de silence sur images naturelles pour le locuteur 6 du corpus "Cardiff". Figures 4.15(a) et 4.15(b) influence de l'intégration (4.27) pour les vues de face et de côté respectivement. Les légendes indiquent la valeur de τ. Figures 4.15(c) et 4.15(d) : influence de la durée minimale des silences détectables pour les vues de face et de côté respectivement. Les légendes indiquent la durée des silences en nombre de trames.

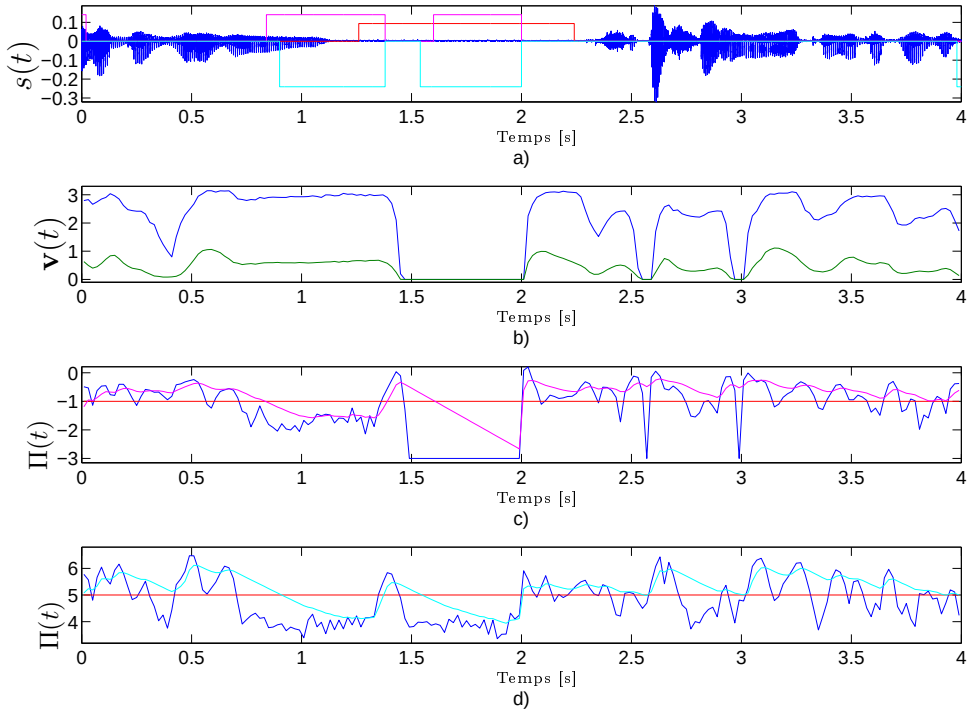


FIG. 4.16 – Exemple de détection des silences. Figure a) : signal acoustique $s(t)$ avec la référence des silences (rouge), les trames détectées *silence* (magenta) par le détecteur de silence du paragraphe 4.3.1 et celles détectées silence (cyan) par celui du paragraphe 4.3.2. Figure b) : évolution de la largeur (vert) et hauteur (bleu) internes des lèvres. Figure c) : logarithmes décimaux des paramètres vidéo instantané tronqué à -3 (4.25) (bleu) et intégré (4.27) avec $\tau = 5$ (magenta). La valeur du seuil est indiquée en rouge. Figure d) : logarithmes décimaux des paramètres vidéo instantané (4.32) (bleu) et intégré (4.34) avec $\tau = 5$ (cyan). La valeur du seuil est indiquée en rouge.

des silences. Il est intéressant de noter que l'on obtient des performances similaires avec la détection visuelle de silence sur images naturelles que celles obtenues avec le détecteur visuel de silence sur les largeur et hauteur internes des lèvres. De même dans des conditions moins contrôlées obtenues à partir du corpus "Cardiff", l'intégration temporelle permet d'améliorer les performances (figure 4.15). La détection des silences par la vue de côté donne des résultats similaires à ceux obtenus par la vue de face, prouvant qu'il est donc possible de faire un détecteur visuel de silence à partir d'une vue de côté du locuteur.

Finalement, la figure 4.16 montre un exemple de l'indexation des silences par les détecteurs de silence visuels des paragraphes 4.3.1 et 4.3.2. Cette section de parole de 4 secondes regroupe toutes les relations possibles entre les données visuelles et le signal de parole : mouvement des lèvres pendant la parole et le silence, pas de mouvements des lèvres pendant la parole et pendant le silence. Nous pouvons voir que les détecteurs de silence intégrés ($\tau = 5$ pour les deux) permettent de

bien détecter les silences. Ils ne peuvent empêcher des fausses détections de silence (entre 0.8s et 1.1s) ou des non détections de silence (entre 2s et 2.2s). Cependant, l'intégration permet d'éviter des fausses détections de silence que produiraient les détecteurs de silence instantanés (courbes bleues des figures 4.16c) et 4.16d)).

4.6 En résumé

Dans ce chapitre, nous avons étendu la détection statistique d'activité de parole purement acoustique développée dans la littérature à une détection bimodale d'activité vocale exploitant le modèle audiovisuel normalisé du chapitre 3. Comme l'ont montré nos expériences, un tel détecteur audiovisuel d'activité vocale est plus performant qu'un détecteur d'activité vocal purement acoustique traditionnel. Ce détecteur bimodal d'activité vocale nécessite une connaissance statistique du bruit, mise à jour au cours du temps pour suivre ses évolutions. Dans le cas de bruit fortement non stationnaire, ce suivi n'est pas aisé. Ainsi, nous avons proposé un détecteur de silence purement visuel fondé sur l'hypothèse de non mouvement des lèvres pendant le silence. Proposé dans un premier temps sur les paramètres de hauteur et largeur internes du contour labial, ce détecteur visuel d'activité vocale a ensuite été proposé sur des images naturelles. Afin de tester ce dernier, nous avons conçu le protocole et l'enregistrement d'une base de données originale qui n'a, pour l'instant, été exploitée que partiellement. Les performances d'un tel détecteur visuel d'activité vocale sont très bonnes et présentent l'avantage de ne pas dépendre du bruit acoustique notamment si celui-ci est fortement non stationnaire.

Troisième partie

Extraction de source de parole audiovisuelle

Introduction

La partie précédente nous a permis de modéliser la bimodalité de la parole sous deux approches différentes que ce soit par un modèle statistique audiovisuel (chapitre 3) ou par un détecteur d'activité vocale audiovisuel ou purement visuel (chapitre 4). Nous allons maintenant exploiter ces modélisations de façon à les inclure dans des algorithmes de séparation de sources, et ceci de deux façons différentes, d'une part comme une alternative aux techniques purement acoustiques actuellement utilisées et d'autre part comme base de nouveaux procédés de séparation.

Pour cela, rappelons brièvement le contexte de la séparation de sources exposée au chapitre 2. Nous disposons de N_o observations, notées $\mathbf{x}(t) = [x_1(t), \cdots, x_{N_o}(t)]^T$, de N_s sources, notées $\mathbf{s}(t) = [s_1(t), \cdots, s_{N_s}(t)]^T$, obtenues à partir d'une fonction de mélange $\mathcal{H}(\cdot)$

$$\mathbf{x}(t) = \mathcal{H}(\mathbf{s}(t)).$$

La séparation de sources consiste à construire uns fonction de séparation $\mathcal{G}(\cdot)$ telle que ses sorties $\mathbf{y}(t)$ soient des estimées des sources $\mathbf{s}(t)$

$$\mathbf{y}(t) = \mathcal{G}(\mathbf{x}(t)).$$

Dans le cadre de mélanges convolutifs, les fonctions de mélange et de séparation sont supposés être linéaires et avec mémoire. On peut alors écrire que

$$\mathbf{x}(t) = H(t) * \mathbf{s}(t) \tag{III.1}$$

et

$$\mathbf{y}(t) = G(t) * \mathbf{x}(t). \tag{III.2}$$

Dans notre étude, nous cherchons à séparer, ou plutôt extraire, une source de parole particulière que nous appellerons source d'intérêt. Arbitrairement, nous considérerons que cette source d'intérêt est par exemple la première notée $s_1(t)$. Pour cela, nous disposons non seulement des observations acoustiques $\mathbf{x}(t)$ mais également d'une observation visuelle additionnelle, $\mathbf{v}_1(t)$, des lèvres du locuteur d'intérêt dont nous cherchons à extraire le signal acoustique correspondant (*cf.* figure 4.17). Pour extraire la seule source $s_1(t)$ des mélanges, nous pouvons nous contenter de déterminer la première ligne de la matrice de séparation $G(t)$ et nous noterons $G_{1,:}(t)$ le vecteur ligne tel que $G_{1,:}(t) = [G_{1,1}(t), \cdots, G_{1,N_o}(t)]$.

Suivant que nous utiliserons le modèle statistique audiovisuel que nous avons construit au chapitre 3, ou que nous exploiterons la détection d'activité vocale multimodale introduite au chapitre 4, nous abordons ce problème d'extraction de sources de parole audiovisuelle soit pour résoudre les indéterminations (permutations et/ou

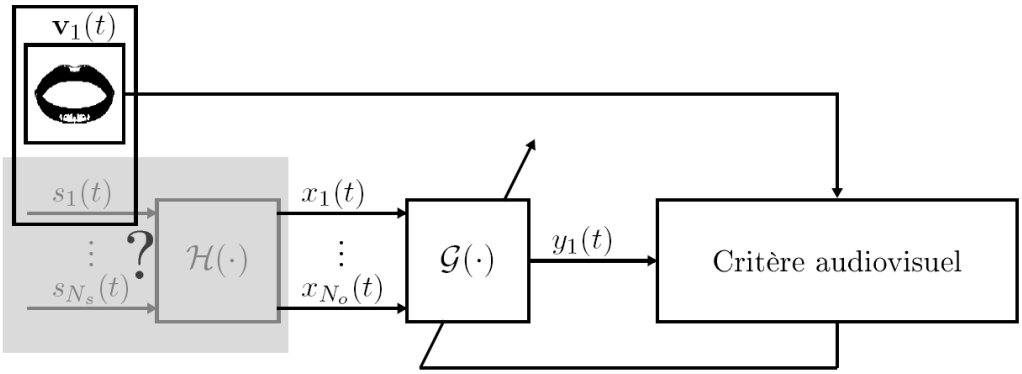


FIG. 4.17 – Extraction d'une source de parole audiovisuelle.

gains) rencontrées par les méthodes de séparation aveugle de source, soit pour estimer directement la ligne $G_{1,:}(t)$ de la matrice de séparation permettant ainsi une extraction directe de la source d'intérêt. Notons que les méthodes que nous proposons permettent d'extraire des observations $\mathbf{x}(t)$ la source pour laquelle nous disposons du signal vidéo. Si nous souhaitons extraire d'autres sources de parole par les méthodes que nous proposons, il est nécessaire de disposer des observations visuelles correspondantes.

Chapitre 5

Extraction par la résolution des indéterminations

Dans ce chapitre, nous présentons deux techniques audiovisuelles appliquées après un algorithme de séparation de sources dans le but d'extraire une source de parole à partir d'observations obtenues par un mélange convolutif de sources. L'idée principale de ces méthodes est d'exploiter la cohérence audiovisuelle de la parole pour résoudre le problème des indéterminations (gain et permutation) rencontrées à chaque fréquence. Pour cela, on s'appuie sur la modélisation des cohérences audiovisuelles soit grâce au modèle audiovisuel proposé au chapitre 3, soit par le détecteur d'activité vocale visuel du chapitre 4.

Après avoir rappelé le principe de séparation de sources dans des mélanges convolutifs et le problème des indéterminations qui en découle, nous introduirons deux nouvelles méthodes audiovisuelles pour les résoudre.

5.1 Position du problème

Nous rappelons dans ce paragraphe les indéterminations dues à la séparation de sources dans le domaine fréquentiel par un critère d'indépendance.

5.1.1 Indéterminations

Dans le cas de mélanges convolutifs, la résolution du problème de séparation de sources de mélange consiste, comme nous venons de le rappeler, à estimer une matrice de filtres $G(t)$. Comme plusieurs auteurs l'ont proposé, voir par exemple [93, 40, 102, 108, 139, 105], nous allons effectuer la séparation dans le domaine fréquentiel où les équations de mélange (III.1) et de séparation (III.2) deviennent respectivement

$$\mathbf{X}(t,f) = H(f)\,\mathbf{S}(t,f) \tag{5.1}$$

et

$$\mathbf{Y}(t,f) = G(f)\,\mathbf{X}(t,f), \tag{5.2}$$

où $G(f)$ et $H(f)$ sont respectivement les réponses en fréquence des fonctions de mélange $H(t)$ et de séparation $G(t)$. Puisque les fonctions de mélange et de séparation sont supposées stationnaires, $H(f)$ et $G(f)$ ne dépendent pas du temps,

tandis que les signaux (*i.e.* sources, observations) peuvent être non stationnaires. $\mathbf{S}(t,f) = [S_1(t,f), \cdots, S_{N_s}(t,f)]^T$, $\mathbf{X}(t,f) = [X_1(t,f), \cdots, X_{N_o}(t,f)]^T$ et $\mathbf{Y}(t,f) = [Y_1(t,f), \cdots, Y_{N_s}(t,f)]^T$ sont respectivement les vecteurs regroupant les transformées de Fourier à court terme (TFCT) des N_s sources, des N_o observations et des N_s estimées des sources à l'instant t et à la fréquence f. Pour estimer la matrice de séparation $G(f)$, nous utilisons la méthode proposée par Pham *et al.* [102] exploitant la non-stationnarité des sources dont nous rappelons ici brièvement le principe de fonctionnement. Cette méthode repose sur le fait que, pour des sources indépendantes, les matrices de densités spectrales de puissance à court terme des sources $\Gamma_{\mathbf{S},\mathbf{S}}(t,f)$ sont diagonales quelles que soient la fréquence f et l'instant t. Les sources étant de plus supposées non-stationnaires, à une fréquence donnée f, les matrices $\Gamma_{\mathbf{S},\mathbf{S}}(t,f)$ évoluent au court du temps. Ainsi, en se limitant à l'ordre deux, Pham *et al.* [102] proposent, pour estimer la matrice de séparation $G(t)$, de diagonaliser conjointement un ensemble de matrices $\{G(f)\Gamma_{\mathbf{X},\mathbf{X}}(t,f)G^+(f)\}_t$ par un algorithme de diagonalisation rapide [96]. Notons que cette méthode nécessite que les spectres d'amplitude des sources évoluent différemment au cours du temps. Cependant, comme toute méthode de séparation de sources fondée sur l'indépendance, ce principe ne permet d'estimer la matrice de séparation qu'à une permutation et un gain près. Dans le cas de séparation dans le domaine fréquentiel, ce problème se rencontre à chaque fréquence f :

$$\forall\, f,\ \exists\, (\Pi(f), \Lambda(f))\ /\ G(f) = \Pi(f)\,\Lambda(f)\,H^{-1}(f), \tag{5.3}$$

où $\Pi(f)$ et $\Lambda(f)$ sont respectivement les matrices de permutation et de gain à la fréquence f. La limitation de cette méthode vient de ce que l'estimation de $G(f)$ est faite à chaque fréquence indépendamment des autres fréquences : ainsi rien n'assure que les matrices $\Pi(f)$ soient les mêmes pour toutes les fréquences f comme illustré à la figure 5.1. Pour obtenir une bonne reconstruction des sources il est alors nécessaire de s'assurer que les permutations et les distorsions soient les mêmes à chaque fréquence f :

$$\forall (f_1, f_2), \quad \begin{cases} \Pi(f_1) = \Pi(f_2), \\ \Lambda(f_1) = \Lambda(f_2). \end{cases}$$

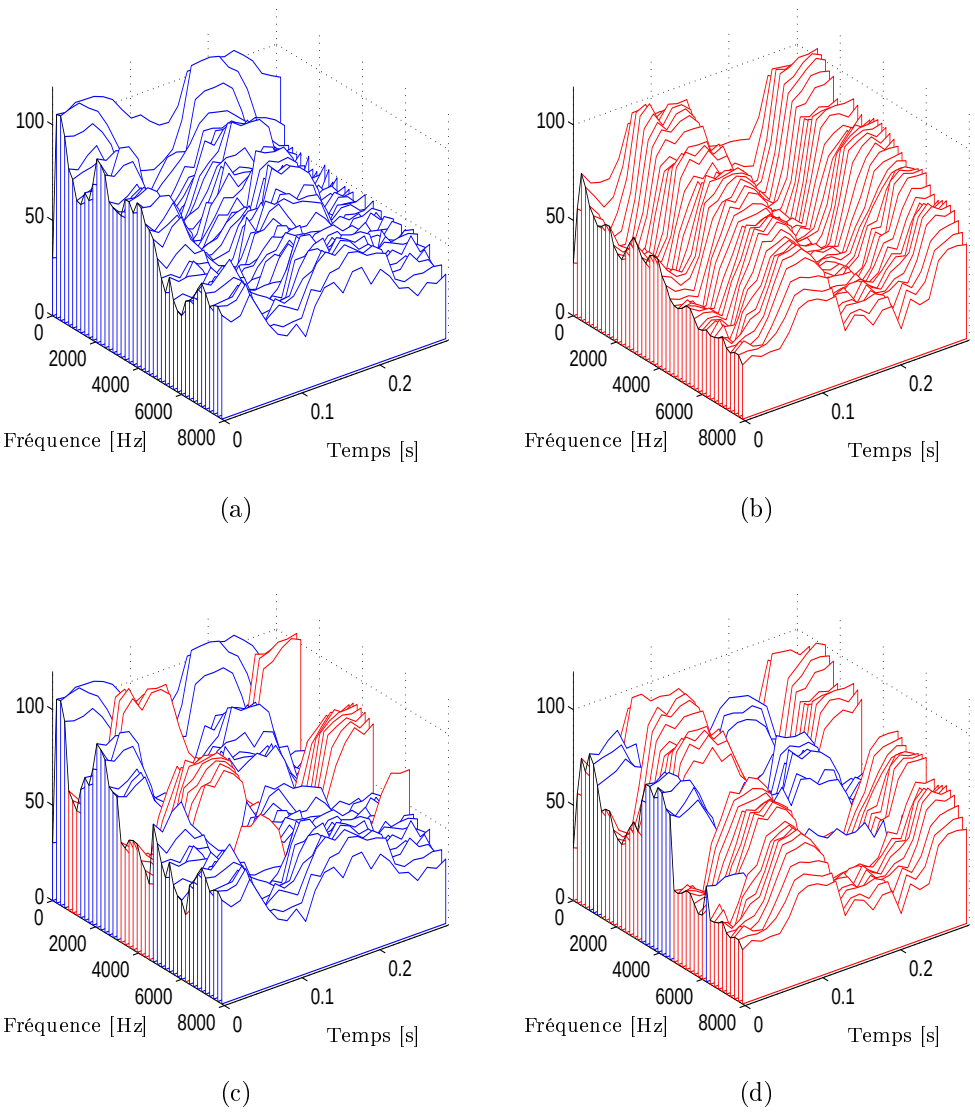
Notons que, puisque l'on cherche uniquement à extraire la première source $s_1(t)$, il n'est alors nécessaire de lever les indéterminations que pour cette source : c'est-à-dire être capable d'assurer

$$\begin{cases} \Pi_{1,1}(f) = 1, & \forall\, f, \qquad (5.4a) \\ \Lambda_{1,1}(f_1) = \Lambda_{1,1}(f_2), & \forall\, (f_1, f_2). \qquad (5.4b) \end{cases}$$

Plusieurs solutions purement audio ont été proposées en faisant des hypothèses sur le processus de mélange ou sur les sources (*cf* chapitre 2). Comme nous allons le voir dans ce chapitre, nous proposons deux nouvelles solutions audiovisuelles pour résoudre ce problème des indéterminations.

5.1.2 Notations

Les sources $S_n(t,f)$ peuvent être vues d'un point de vue mathématique comme des matrices à trois dimensions : (indice de la $n^{\text{ème}}$ source $\times$ temps t $\times$ fréquence f).



(a) (b)

(c) (d)

FIG. 5.1 – Problème des permutations pour la séparation fréquentielle : densité spectrale de puissance à court terme $\Gamma_{ss}(t, f)$ de deux sources indépendantes ([UPUP] 5.1(a) et [APAPA] 5.1(b)) et des deux sources estimées avec des permutations nuisibles (5.1(c) et 5.1(d)).

De façon à lever toute ambiguïté sur les notations vectorielles, nous noterons par ":" la dimension qui est vectorisée. Ainsi, par exemple $\mathbf{S}_n(t,:)$ est le vecteur colonne qui regroupe les coefficients de la TFCT de la $n^{\text{ème}}$ source à l'instant t : $\mathbf{S}_n(t,:) = [S_n(t,f_1),\cdots,S_n(t,f_{N_f})]^T$.

Soit $\mathbf{Y}_:(t,f) = [Y_1(t,f),\cdots,Y_{N_s}(t,f)]^T$ le vecteur colonne des N_s estimées des sources à l'instant t et la fréquence f. En faisant l'hypothèse que l'algorithme de diagonalisation conjointe fournit une matrice $G(f)$ séparante, ce vecteur est donc égal à $\mathbf{S}_:(t,f) = [S_1(t,f),\cdots,S_{N_s}(t,f)]^T$ à une permutation $\Pi(f)$ et un gain près $\Lambda(f)$:

$$\mathbf{Y}_:(t,f) = \Pi(f)\,\Lambda(f)\,\mathbf{S}_:(t,f).$$

Finalement, dans un souci de simplicité, nous noterons Π_f (resp. Λ_f) l'ensemble des matrices de permutation (resp. diagonales de gain) $\{\Pi(f)\}_f$ (resp. $\{\Lambda(f)\}_f$).

5.2 De la cohérence audiovisuelle...

Dans ce paragraphe, nous introduisons, pour lever les indéterminations, notre nouvelle approche qui exploite la cohérence audiovisuelle d'un signal de parole grâce au modèle audiovisuel introduit au chapitre 3. Nous expliquons tout d'abord le principe pour régulariser les permutations. Puis nous présenterons la résolution du problème des distorsions (gains) et finalement nous présenterons comment nous exploitons ces deux régularisations ensemble [111].

5.2.1 Indétermination de permutation

Supposons dans un premier temps que $\forall\, f,\ \Lambda(f) = I_{N_s}$. Régulariser alors le problème des permutations produites par la séparation des sources dans le domaine fréquentiel (*i.e.* satisfaire (5.4a)), consiste à chercher l'ensemble $\widehat{\mathcal{P}}_f$ des matrices de permutation tel qu'à chaque fréquence f

$$Y_1\Big(t,f\,\Big|\widehat{\mathcal{P}}(f)\Big) = S_1(t,f), \tag{5.5}$$

où $Y_1\big(t,f|\mathcal{P}(f)\big)$ est l'estimée du coefficient de la TFCT de $s_1(t)$, pour l'instant t et la fréquence f, à la permutation $\mathcal{P}(f)$ près. Notons qu'à chaque fréquence f la matrice de permutation ne dépend pas du temps t puisque la matrice de séparation $G(f)$ est indépendante du temps. Ainsi, nous avons

$$Y_1\big(t,f\,|\mathcal{P}(f)\big) = \big(\mathcal{P}(f)\,\mathbf{Y}_:(t,f)\big)_1 \tag{5.6}$$

où $(\mathbf{z})_k$ est la $k^{\text{ème}}$ composante du vecteur $\mathbf{z}$. Par simplicité, nous notons $\mathbf{Y}_1\big(t,:|\mathcal{P}_f\big)$ le vecteur dont les composantes sont définies par $Y_1\big(t,f\,|\mathcal{P}(f)\big)$:

$$\mathbf{Y}_1\big(t,:|\mathcal{P}_f\big) = \Big[Y_1\big(t,f_1\,|\mathcal{P}(f_1)\big),\cdots,Y_1\big(t,f_{N_f}\,\big|\mathcal{P}(f_{N_f})\big)\Big]^T.$$

Pour estimer $\mathcal{P}_f$, nous proposons de minimiser le critère audiovisuel $j_1^{AV}\big(t|\mathcal{P}_f\big)$ entre le spectre d'amplitude acoustique de la première sortie $y_1(t)$ et l'information visuelle $\mathbf{v}_1(t)$:

$$\widehat{\mathcal{P}}_f = \arg\min_{\mathcal{P}_f} j_1^{AV}\big(t\,|\mathcal{P}_f\big) \tag{5.7}$$

avec

$$j_1^{AV}\big(t\,|\mathcal{P}_f\big) = -\ln\Big[p_{AV}\Big(\mathbf{a}_{Y_1}\big(t,:|\mathcal{P}_f\big),\mathbf{v}_1(t)\Big)\Big], \tag{5.8}$$

où $p_{AV}(\cdot)$ est la densité de probabilité du modèle audiovisuel défini au chapitre 3 par l'équation (3.18) et $\mathbf{a}_{Y_1}\big(t,:|\mathcal{P}_f\big)$ est le logarithme, composante à composante, du module de $\mathbf{Y}_1\big(t,:|\mathcal{P}_f\big)$:

$$\mathbf{a}_{Y_1}\big(t,:|\mathcal{P}_f\big) = \ln\Big|\mathbf{Y}_1\big(t,:|\mathcal{P}_f\big)\Big|.$$

Notons que, comme le critère (5.8) ne porte que sur l'estimation de la première source $y_1(t)$, il n'assure au mieux que $\big(\widehat{\mathcal{P}}(f)\Pi(f)\big)_{1,1} = 1$ pour toutes les fréquences f, laissant les autres termes de $\big(\widehat{\mathcal{P}}(f)\Pi(f)\big)$ non-spécifiés. En d'autres termes, ceci signifie que, au mieux, la méthode proposée assure seulement que les composantes de $\mathbf{Y}_1\big(t,:|\widehat{\mathcal{P}}_f\big)$ correspondent effectivement à celles de la source d'intérêt (*i.e.* que (5.5) soit vérifiée) sans aucune contrainte sur les autres estimations des sources qui peuvent alors contenir encore des permutations. Mais, dans notre problématique, ceci n'est pas un réel problème puisque nous voulons seulement extraire la première source $s_1(t)$ à partir de $\mathbf{v}_1(t)$. On rappelle qu'extraire d'autres sources de parole avec notre méthode requerrait des observations vidéo supplémentaires sur ces sources à extraire.

Dans [120, 118], il a été montré que pour exploiter correctement la cohérence audiovisuelle, il est nécessaire de faire une intégration temporelle du modèle statistique audiovisuel (3.18) puisque la cohérence audiovisuelle est principalement exprimée dans la dynamique temporelle de la parole. Ainsi, pour améliorer le critère (5.8), nous introduisons la possibilité de lisser les probabilités au cours du temps. Dans ce but, nous supposons que les valeurs des caractéristiques audio et vidéo pour différents instants consécutifs sont indépendantes et nous définissons un critère audiovisuel intégré par

$$J_1^{AV}\big(\mathcal{P}_f\big) = \sum_{t=0}^{T-1} j_1^{AV}\big(t\,|\mathcal{P}_f\big). \tag{5.9}$$

Comme il y a $\big(N_s!\big)^{N_f}$ matrices de permutation possible (si le nombre de sources est N_s et N_f le nombre de fréquences de calcul des TFCT), il n'est pas possible de faire une recherche exhaustive du fait du coût calculatoire que cela implique. Pour simplifier la présentation de l'algorithme que nous proposons pour la minimisation du critère (5.9), nous ne présentons le principe que pour deux sources, mais ce principe est facilement généralisable aux cas où plus de deux sources rentrent en jeu. Tout d'abord, nous utilisons un algorithme dichotomique dans lequel nous simplifions le critère (5.9) par une marginalisation de la probabilité audiovisuelle $p_{AV}(\cdot,\cdot)$ vis-à-vis de sous-ensembles de fréquences consécutives. Ainsi, soient $\mathcal{F}$ un sous-ensemble arbitraire de fréquences f_j et $p_{AV}^{\mathcal{F}}(\cdot,\cdot)$ la densité de probabilité audiovisuelle marginale vis-à-vis du sous-ensemble $\mathcal{F}$:

$$p_{AV}^{\mathcal{F}}\big(\mathbf{a}(t),\mathbf{v}(t)\big) = \int_{\forall\, f_j\notin\mathcal{F}}\cdots\int p_{AV}\big(\mathbf{a}(t),\mathbf{v}(t)\big)\,da(t,f_j). \tag{5.10}$$

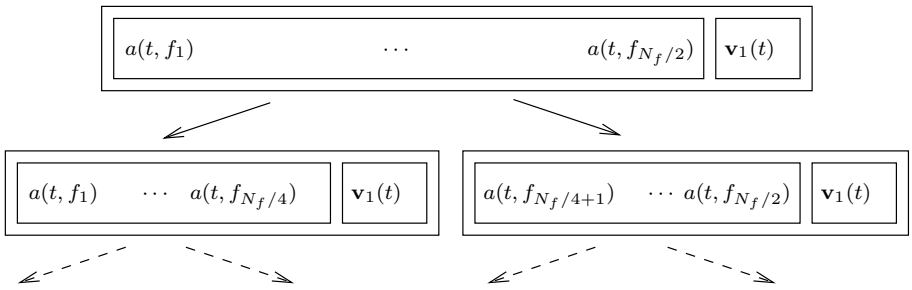


FIG. 5.2 – Algorithme marginal audiovisuel : à chaque étape, nous testons avec le critère audiovisuel marginal (5.11) s'il est nécessaire d'effectuer la permutation d'un ensemble de coefficients audio dont le nombre est divisé par deux à chaque fois.

On définit alors la forme marginale de (5.9) par

$$J_1^{AV}\big(\mathcal{P}_f, \mathcal{F}\big) = \sum_{t=0}^{T-1} j_1^{AV}\big(t\,|\mathcal{P}_f, \mathcal{F}\big) \tag{5.11}$$

où

$$j_1^{AV}\big(t\,|\mathcal{P}_f, \mathcal{F}\big) = -\ln\Big[p_{AV}^{\mathcal{F}}\Big(\mathbf{a}_{Y_1}\big(t,:|\mathcal{P}_f\big), \mathbf{v}_1(t)\Big)\Big]. \tag{5.12}$$

Nous allons maintenant exploiter cette simplification en utilisant l'algorithme dichotomique descendant suivant, que nous appelons algorithme marginal audiovisuel (*cf.* figure 5.2)

1. Tout d'abord, on teste la permutation globale de tous les paramètres audio (*i.e.* $\mathcal{F} = \{f_1, \cdots, f_{N_f/2}\}$), entre les deux sources estimées, qui minimise $J_1^{AV}\big(\cdot, \mathcal{F}\big)$

$$J_1^{AV}(\mathcal{J}_{f\in\mathcal{F}}, \mathcal{F}) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}_{f\in\mathcal{F}}, \mathcal{F})$$

   où $\mathcal{I}_f$ est l'ensemble de matrices identités et $\mathcal{J}_f$ est l'ensemble de matrices unitaires anti-diagonales[1]. H_r signifie qu'il ne faut pas faire de permutation puisque le vecteur $\mathbf{a}_{Y_1}\big(t,:|\mathcal{I}_{f\in\mathcal{F}}\big) = \mathbf{a}_{Y_1}(t,:)$ est plus cohérent avec la vidéo $\mathbf{v}_1(t)$ que le vecteur[2] $\mathbf{a}_{Y_1}\big(t,:|\mathcal{J}_{f\in\mathcal{F}}\big) = \mathbf{a}_{Y_2}(t,;)$ qui correspond à la permutation de tous les coefficients. H_p signifie qu'il faut effectuer la permutation, aux fréquences incluses dans $\mathcal{F}$, entre les coefficients des deux sources estimées.

2. On affine ensuite l'estimation de l'ensemble des matrices de permutations en testant séparément par le critère marginal (5.11).

[1] Nous rappelons que dans un souci de simplicité, nous ne présentons notre principe que dans le cas de deux sources. Il n'y a alors que deux possibilités pour les matrices de permutation : $\mathcal{I}_f$ qui correspond à aucune permutation et $\mathcal{J}_f$ qui permute les coefficients entre les deux sources estimées.

[2] Remarquons que si $\mathbf{z}$ un vecteur colonne et Π une matrice de permutation, alors on a $\Pi \ln \mathbf{z} = \ln\big(\Pi\,\mathbf{z}\big)$, où ln est le logarithme composante à composante. Ceci signifie qu'il est équivalent d'effectuer une permutation des composantes d'un vecteur puis de prendre le logarithme du résultat ou bien d'effectuer la permutation directement sur le logarithme des composantes du même vecteur.

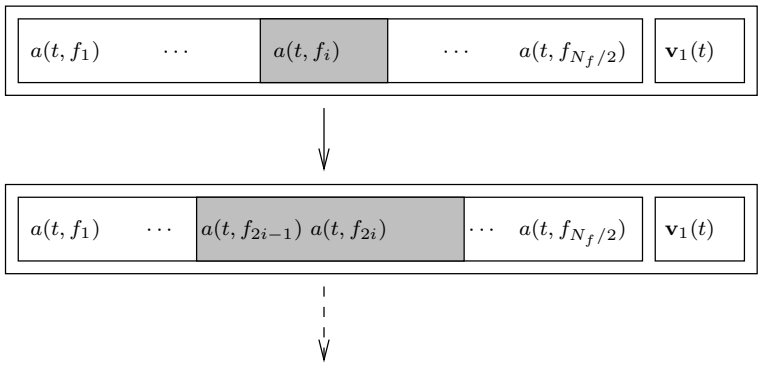


FIG. 5.3 – Algorithme audiovisuel conjoint : à chaque étape, on teste grâce au critère conjoint (5.9), s'il faut effectuer la permutation d'un ensemble de coefficients audio dont le nombre est multiplié par deux à chaque fois.

- une permutation de l'ensemble de la première moitié des fréquences correspondantes aux $N_f/4$ plus petites fréquences, $\mathcal{F}_1 = \{f_1, \cdots, f_{N_f/4}\}$:

$$J_1^{AV}(\mathcal{J}_{f\in\mathcal{F}_1}, \mathcal{F}_1) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}_{f\in\mathcal{F}_1}, \mathcal{F}_1),$$

- une permutation de l'ensemble de la seconde moitié des fréquences correspondantes aux $N_f/4$ plus grandes fréquences, $\mathcal{F}_2 = \{f_{N_f/4+1}, \cdots, f_{N_f/2}\}$:

$$J_1^{AV}(\mathcal{J}_{f\in\mathcal{F}_2}, \mathcal{F}_1) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}_{f\in\mathcal{F}_2}, \mathcal{F}_2).$$

3. On continue ainsi de suite avec ce principe dichotomique sur tous les sous-ensembles de fréquences dont on aura divisé la taille par un facteur 2.

Ensuite, le principe de cet algorithme consiste à rechercher, à chacune des étapes, quel est le sous-vecteur des coefficients audio entre $[a_{Y_1}(t,f_i), \cdots, a_{Y_1}(t,f_j)]^T$ et $[a_{Y_2}(t,f_i), \cdots, a_{Y_2}(t,f_j)]^T$ le plus cohérent avec le vecteur vidéo $\mathbf{v}_1(t)$ et ce au sens du critère audiovisuel marginal (5.11).

Ce premier algorithme, utilisant le critère audiovisuel marginal (5.11), donne une bonne estimation de l'ensemble des permutations $\mathcal{P}_f$, mais nous observons que ce résultat peut être amélioré. Ceci n'est pas surprenant puisque la densité de probabilité audiovisuelle marginale (5.10) ne tient pas compte de toutes les cohérences audiovisuelles. Nous raffinons alors l'estimation de $\mathcal{P}_f$ grâce à l'algorithme récursif ascendant suivant que nous appelons algorithme audiovisuel conjoint (*cf.* figure 5.3) :

1. Pour chaque fréquence f_i, $1 \leq i \leq N_f/2$, tester avec (5.9) l'ensemble des permutations $\mathcal{P}_{f_i}$, entre les coefficients des deux sources estimées, qui permute ceux-ci à la fréquence f_i laissant inchangé les coefficients des autres fréquences f_j, $j \neq i$

$$J_1^{AV}\big(\mathcal{P}_{f_i}\big) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}).$$

En d'autres termes, on cherche entre les vecteurs $[a_{Y_1}(t,f_1),\cdots,a_{Y_1}(t,f_{N_f})]^T$ et $[a_{Y_1}(t,f_1),\cdots,a_{Y_1}(t,f_{i-1}),a_{Y_2}(t,f_i),a_{Y_1}(t,f_{i+1}),\cdots,a_{Y_1}(t,f_{N_f})]^T$ lequel est le plus cohérent avec le vecteur vidéo $\mathbf{v}_1(t)$ au sens du critère (5.9).

2. Ensuite, pour chaque couple de fréquences $\big(f_{2i-1},f_{2i}\big)$, $1\leq i\leq N_f/4$, on teste de façon similaire l'ensemble des permutations $\mathcal{P}_{2i-1,2i}$, entre les coefficients des deux sources estimées, qui permute ceux-ci au couple de fréquences $\big(f_{2i-1},f_{2i}\big)$, laissant inchangé les coefficients autres aux fréquences

$$J_1^{AV}(\mathcal{P}_{2i-1,2i}) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}).$$

En d'autres termes, on cherche entre les vecteurs $[a_{Y_1}(t,f_1),\cdots,a_{Y_1}(t,f_{N_f})]^T$ et $[a_{Y_1}(t,f_1),\cdots,a_{Y_1}(t,f_{2i-2}),a_{Y_2}(t,f_{2i-1}),a_{Y_2}(t,f_{2i}),a_{Y_1}(t,f_{2i+1}),\cdots,a_{Y_1}(t,f_{N_f})]^T$ lequel est le plus cohérent avec le vecteur vidéo $\mathbf{v}_1(t)$ au sens du critère (5.9).

3. On continue avec ce principe dichotomique jusqu'à tester l'ensemble des permutations $\mathcal{P}_{1,\cdots,N_f/2}$ qui permute les coefficients pour l'ensemble des fréquences $\{f_1,\cdots,f_{N_f/2}\}$

$$J_1^{AV}\big(\mathcal{P}_{1,\cdots,N_f/2}\big) \underset{H_r}{\overset{H_p}{\lessgtr}} J_1^{AV}(\mathcal{I}).$$

4. On boucle à l'étape 1 si nécessaire.

La succession des deux algorithmes s'est révélée relativement efficace comme on le verra à la section 5.4. Malheureusement, un problème délicat apparaît quand on utilise les critères (5.9) et (5.11) : le problème du facteur d'échelle. En effet, l'ensemble des matrices de gain Λ_f peut influencer de façon dramatique les valeurs des densités de probabilité dans les équations (5.8) et (5.10). Ainsi, ce facteur d'échelle inconnu doit être estimé de façon à améliorer la qualité de la reconstruction des sources. C'est pourquoi nous présentons maintenant une méthode pour estimer l'ensemble des facteurs d'amplitude rencontrés à chaque fréquence.

5.2.2 Estimation des facteurs d'amplitude

La régularisation du facteur d'amplitude (5.4b) pour un problème de séparation de sources dans le domaine fréquentiel consiste à estimer un ensemble de gain $\widehat{\mathcal{L}}_f$ qui vérifie

$$Y_1\Big(t,f\big|\widehat{\mathcal{L}}(f)\Big)=S_1(t,f), \tag{5.13}$$

où $Y_1\Big(t,f\big|\mathcal{L}(f)\Big)$ est l'estimée du coefficient de la TFCT de $s_1(t)$, à l'instant t et à la fréquence f, au gain $\mathcal{L}(f)$ près :

$$Y_1\Big(t,f\big|\mathcal{L}(f)\Big)=\big(\mathcal{L}(f)\,\mathbf{Y}_{:}(t,f)\big)_1.$$

Notons que, puisque la matrice de séparation $G(f)$ ne dépend pas du temps, l'ensemble des matrices de gain $\mathcal{L}_f$ ne dépend pas non plus du temps. Ainsi, nous avons

$$Y_1\Big(t,f\big|\mathcal{L}(f)\Big)=\mathcal{L}_{1,1}(f)\,Y_1(t,f)$$

puisque $\mathcal{L}(f)$ est une matrice diagonale. Notons que, même si l'estimation optimale de $\widehat{\mathcal{L}}(f)$ était $\Lambda^{-1}(f)$, puisque nous ne cherchons qu'à extraire la première source, il suffit d'estimer le coefficient $\widehat{\mathcal{L}}_{1,1}(f)$. Ainsi, chercher à régulariser le facteur d'amplitude consiste à chercher, à chaque fréquence f, le coefficient $\widehat{\mathcal{L}}_{1,1}(f)$ tel que

$$\widehat{\mathcal{L}}_{1,1}(f)\, Y_1(t,f) = S_1(t,f). \tag{5.14}$$

Pour estimer $\widehat{\mathcal{L}}_{1,1}(f)$, nous proposons d'exploiter le modèle purement audio obtenu par marginalisation du modèle audiovisuel (3.18) du chapitre 3 que nous rappelons ici

$$p_{AV}(\mathbf{a}_1(t,:),\mathbf{v}_1(t)) = \sum_{i=1}^{N_{AV}} \omega_i^{AV}\, p_G\big(\mathbf{v}_1(t)\,\big|\mu_i^V,\Sigma_i^V\big)\; p_{LR}\big(\mathbf{a}_1(t,:)\,\big|\Gamma_i^A\big)\,,$$

où $\mathbf{a}_1(t,:) = \ln|\mathbf{S}_1(t,:)|$ est le logarithme du module des coefficients de la TFCT. La marginalisation de ce modèle vis-à-vis des paramètres vidéo donne le modèle purement audio suivant

$$p_A(\mathbf{a}_1(t,:)) = \sum_{i=1}^{N_A} \omega_i^A\, p_{LR}\left(\mathbf{a}_1(t,:)|\Gamma_i^A\right),$$

où ω_i^A est le poids du $i^{\text{ème}}$ noyau audio dérivé du noyau audiovisuel correspondant : $\omega_i^A = \omega_i^{AV}$ et $N_A = N_{AV}$ est le nombre de noyaux audio. Notons qu'une distribution de LogRayleigh sur $\mathbf{a}_1(t,:)$ est équivalente à une distribution normale complexe circulaire sur $\mathbf{S}_1(t,:)$. Ainsi le modèle suivi par $\mathbf{S}_1(t,;)$ est donné par

$$p_A(\mathbf{S}_1(t,:)) = \sum_{i=1}^{N_A} \omega_i^A\, p_G\left(\mathbf{S}_1(t,:)|0,2\Gamma_i^A\right). \tag{5.15}$$

Puisque la variance de $Y_1\big(t,f\big|\widehat{\mathcal{L}}(f)\big)$ vérifie

$$\mathrm{Var}\Big[Y_1\big(t,f\big|\widehat{\mathcal{L}}(f)\big)\Big] = \Big(\widehat{\mathcal{L}}_{1,1}(f)\Big)^2\,\mathrm{Var}\big[Y_1(t,f)\big],$$

où Var[·] est l'opérateur variance. Ainsi, vérifier (5.14) implique que

$$\Big(\widehat{\mathcal{L}}_{1,1}(f)\Big)^2\,\mathrm{Var}\big[Y_1(t,f)\big] = \mathrm{Var}\big[S_1(t,f)\big]. \tag{5.16}$$

Or la variance des coefficients de la TFCT de la première source est donnée par le modèle audio (5.15) :

$$\mathrm{Var}\big[S_1(t,f)\big] = \sum_{i=1}^{N_A} \omega_i^A\big[2\Gamma_i^A(f)\big]. \tag{5.17}$$

En reportant ce résultat dans (5.16), nous proposons d'estimer $\widehat{\Lambda}_{1,1}(f)$ par

$$\widehat{\mathcal{L}}_{1,1}(f) = \sqrt{\frac{\sum_{i=1}^{N_A}\omega_i^A\big(2\Gamma_i^A(f)\big)}{\mathrm{Var}\big[Y_1(t,f)\big]}}, \tag{5.18}$$

où $\mathrm{Var}\big[Y_1(t,f)\big]$ est estimée par l'estimateur classique de la variance donné par

$$\mathrm{Var}\big[Y_1(t,f)\big] = \frac{1}{T-1}\sum_{t=1}^{T}\left|Y_1(t,f)\right|^2. \tag{5.19}$$

Cependant, cette idée ne peut fonctionner que si la variance des signaux acoustiques observés est égale à celle obtenue à partir du modèle purement audio (5.15). Mais ceci peut ne pas être vérifié puisque la variance donnée par le modèle est une variance moyenne correspondant à l'ensemble des sons prononcés lors de l'apprentissage. Ainsi, cette variance peut être différente de la variance d'une section particulière de la parole notamment lorsque la section de parole que nous voulons séparer est petite devant celle utilisée lors de l'apprentissage et que des sons n'ont pas été prononcés. Pour surmonter cette difficulté, nous proposons d'utiliser le modèle audiovisuel : nous allons ici exploiter la redondance de la modalité visuelle vis-à-vis de la modalité auditive. Pour cela, remplaçons dans (5.16) $\mathrm{Var}\big[S_1(t,f)\big]$ par $\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big]$, avec

$$\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big] = \int_{S_1(t,f)}\left|S_1(t,f)\right|^2 p_{A|V}\big(S_1(t,f)|\mathbf{v}_1(t)\big)dS_1(t,f).$$

Or d'après la règle de Bayes, on a

$$p_{A|V}\big(S_1(t,f)|\mathbf{v}_1(t)\big) = \frac{p_{AV}\big(S_1(t,f),\mathbf{v}_1(t)\big)}{p_V\big((\mathbf{v}_1(t)\big)},$$

où $p_V\big(\mathbf{v}_1(t)\big)$ est le modèle visuel marginal obtenu à partir du modèle audiovisuel par marginalisation par rapport aux coefficients audio :

$$p_V\big(\mathbf{v}_1(t)\big) = \sum_{i=1}^{N_V}\omega_i^V\, p_G\big(\mathbf{v}_1(t)|\mu_i^V,\Sigma_i^V\big)$$

avec $N_V = N_{AV}$ est le nombre de noyaux, et $\omega_i^V = \omega_i^{AV}$ est le poids *a priori* du $i^{\text{ème}}$ noyau. On obtient ainsi

$$\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big] = \sum_{i=1}^{N_{AV}}\frac{\omega_i^{AV}\, p_G\big(\mathbf{v}(t)|\mu_i^V,\Sigma_i^V\big)}{\sum_{j=1}^{N_V}\omega_j^V\, p_G\big(\mathbf{v}(t)|\mu_j^V,\Sigma_j^V\big)}\,2\Gamma_i^A(f). \tag{5.20}$$

Ainsi, en tenant compte du fait que $\omega_i^{AV} = \omega_i^V$ par construction, cette expression se réécrit

$$\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big] = \sum_{i=1}^{N_{AV}} p\big(i|\mathbf{v}_1(t)\big)\,2\Gamma_i^A(f) \tag{5.21}$$

qui fait intervenir $p\big(i|\mathbf{v}_1(t)\big)$ qui est la probabilité *a posteriori* visuelle que le $i^{\text{ème}}$ noyau ait généré le son à l'instant t.

Si l'on dispose de plusieurs observations, en considérant que ces observations sont indépendantes les unes des autres, l'expression (5.21) s'exprime

$$\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big] = \sum_{i=1}^{N_{AV}}\left(\frac{1}{T}\sum_{t=1}^{T} p\big(i|\mathbf{v}_1(t)\big)\right)2\Gamma_i^A(f). \tag{5.22}$$

Finalement en reportant ce résultat dans (5.16) où $\mathrm{Var}\big[S_1(t,f)\big]$ est remplacé par $\mathrm{Var}\big[S_1(t,f)|\mathbf{v}_1(t)\big]$, on obtient

$$\widehat{\mathcal{L}}_{1,1}(f) = \sqrt{\frac{\sum_{i=1}^{N_A}\Big(\frac{1}{T}\sum_{t=1}^{T} p\big(i|\mathbf{v}_1(t)\big)\Big)\big(2\Gamma_i^A(f)\big)}{\mathrm{Var}\big[Y_1(t,f)\big]}}. \tag{5.23}$$

Si l'on compare ce résultat avec celui de la formule (5.18), on constate que les poids *a priori* ω_i^A du modèle audio ont été remplacés par la moyenne temporelle de la probabilité *a posteriori* des noyaux connaissant les observations visuelles. Ces dernières peuvent s'apparenter à des poids *a posteriori* des noyaux.

Nous expliquons maintenant comment utiliser simultanément l'estimation de l'ensemble $\mathcal{P}_f$ et $\mathcal{L}_f$ de façon à régulariser à la fois les gains et les permutations.

5.2.3 Algorithme final

Initialement [109], nous avions proposé d'estimer l'ensemble des facteurs d'amplitude Λ_f après avoir régularisé les permutations par les principes audiovisuels du paragraphe 5.2.1. Mais comme nous l'avons expliqué précédemment, l'ensemble Λ_f des facteurs d'amplitude inconnus peut dramatiquement changer la valeur des probabilités (5.8) et (5.10). Ainsi, pour avoir de meilleures performances, nous proposons un algorithme qui combine les estimations des facteurs d'amplitude et des permutations : à chaque étape des algorithmes audiovisuel marginaux et conjoints, si H_p est choisi (*i.e.* si l'on détecte qu'après permutation le nouveau vecteur des coefficients audio est plus cohérent avec la vidéo $\mathbf{v}_1(t)$), alors nous réestimons les facteurs d'amplitude grâce au processus audiovisuel (5.23). Si nécessaire, il est possible de boucler plusieurs fois l'algorithme conjoint audiovisuel que nous proposons.

Ainsi, cet algorithme estime $\widehat{\mathcal{L}_f}$ et $\widehat{\mathcal{P}_f}$, puis les sources estimées $\hat{\mathbf{S}}_{:}(t,f)$ sont obtenues par

$$\hat{\mathbf{S}}_{:}(t,f) = \widehat{\mathcal{L}}(f)\,\widehat{\mathcal{P}}(f)\,\mathbf{Y}_{:}(t,f). \tag{5.24}$$

Finalement, la source d'intérêt $s_1(t)$ est reconstruite par transformée de Fourier inverse à court-terme de $\hat{S}_1(t,f)$.

Notons que, bien que nous n'ayons présenté ici que le cas d'une seule source d'intérêt, il est très facile d'étendre les critères (5.9), (5.11) et (5.23) à un plus grand nombre de sources d'intérêt, si les informations visuelles correspondantes sont disponibles. L'estimation finale des sources reste, quant à elle, toujours donnée par (5.24).

L'algorithme que nous venons de proposer est fondé sur la possibilité d'estimer l'ensemble des permutations Π_f grâce au modèle audiovisuel du chapitre 3. Bien qu'efficace, comme nous le verrons un peu plus tard au paragraphe 5.4 portant sur les expérimentations, ce principe requiert au préalable l'apprentissage du modèle audiovisuel. C'est pourquoi, nous allons présenter dans le paragraphe suivant une autre technique où le modèle audiovisuel est remplacé par le détecteur de silence purement visuel du chapitre 3.

5.3 . . . à la parcimonie de la parole

Une alternative à la solution du problème des permutations donnée dans le paragraphe précédent est d'exploiter la détection des moments de silence de la source d'intérêt $s_1(t)$.

A chaque fréquence f, la fonction de séparation (*cf.* paragraphe 5.1.1) est une matrice de séparation $G(f)$ qui permet de diagonaliser à la fréquence f l'ensemble des matrices $\Gamma_{\mathbf{XX}}(t,f)$ des densités spectrales de puissance à court terme des observations $\mathbf{X}(t)$:

$$\Gamma_{\mathbf{YY}}(t,f) = G(f)\,\Gamma_{\mathbf{XX}}(t,f)\,G^{+}(f), \tag{5.25}$$

où $\Gamma_{\mathbf{YY}}(t,f)$ est la matrice des densités spectrales de puissance des sorties de la séparation. Nous rappelons que les sorties $\mathbf{y}(t)$ étant indépendantes, les matrices $\Gamma_{\mathbf{YY}}(t,f)$ sont diagonales pour tous les instants t et toutes les fréquences f. Le logarithme du $k^{\text{ème}}$ terme de la diagonale est appelé profil par Pham *et al.* [102] :

$$E(t,f;k) = \ln\bigl(\Gamma_{\mathbf{YY}}(t,f)\bigr)_{k,k}.$$

Son évolution au cours du temps correspond à la variation de l'énergie spectrale de la $k^{\text{ème}}$ source à la fréquence f au cours du temps t.

Soit $\mathcal{T}$ l'ensemble de tous les indices temporels t. Supposons maintenant qu'un oracle, qui nous est donné par le détecteur d'activité vocale purement visuel du chapitre 4, nous donne l'ensemble $\mathcal{T}_1$ des indices temporels pendant lesquels la source $s_1(t)$, pour laquelle nous cherchons à régulariser les permutations, est absente du mélange. En d'autres termes, le détecteur d'activité vocale visuel permet de connaître les intervalles de temps $\mathcal{T}_1 \subset \mathcal{T}$ durant lesquels le locuteur, pour lequel nous disposons de l'information visuelle $\mathbf{v}_1(t)$, ne parle pas. La densité spectrale de puissance de la source $s_1(t)$ doit donc être nulle durant ces périodes de temps et le profil $E(t,f;\cdot)$, pour $t \in \mathcal{T}_1$, correspondant à l'estimation de $s_1(t)$, doit nécessairement tendre vers moins l'infini :

$$t \in \mathcal{T}_1 \implies s_1(t) = 0 \implies \forall f,\ \exists\, k \ /\ E(t,f;k) \to -\infty. \tag{5.26}$$

Donc pour estimer, à la fréquence f, la permutation $\widehat{\mathcal{P}}(f)$ telle que $Y_1\bigl(t,f\bigl|\widehat{\mathcal{P}}(f)\bigr) = S_1(t,f)$, nous proposons de rechercher parmi les N_s profils possibles celui qui est le plus petit en moyenne pendant $\mathcal{T}_1$:

$$\forall f, \quad \widehat{\mathcal{P}}(f) = \arg\min_{\mathcal{P}(f)} \frac{1}{|\mathcal{T}_1|} \sum_{t\in\mathcal{T}_1} \Bigl(\mathcal{P}(f)\,\mathbf{E}(t,f;:)\Bigr)_1, \tag{5.27}$$

où $\mathbf{E}(t,f;:) = [E(t,f;1),\cdots,E(t,f;N_s)]^T$ est le vecteur dont les composantes sont les N_s profils moyens et $|\mathcal{T}_1|$ est le cardinal de l'ensemble $\mathcal{T}_1$. Cependant, puisque chaque source $s_i(t)$ ne peut être estimée qu'à un gain près, dû au facteur d'amplitude $\Lambda_{i,i}(f)$, les profils moyens sont définis à une constante additive près

$$\forall k, \quad E(t,f;k) = \ln\Bigl(\Pi(f)\Lambda(f)\Gamma_{\mathbf{S},\mathbf{S}}(t,f)\Lambda^{+}(f)\Pi^{+}(f)\Bigr)_{k,k}$$

ainsi

$$\forall k,\ \exists j \ /\ E(t,f;k) = \ln\bigl(\Gamma_{\mathbf{S},\mathbf{S}}(t,f)\bigr)_{j,j} + 2\ln\Lambda_{j,j}(f).$$

Pour s'affranchir de cette constante et donc du facteur d'amplitude, nous définissons à chaque fréquence f le profil moyen centré $E_{\mathcal{T}_1}(f;k)$ calculé sur l'ensemble $\mathcal{T}_1$, ensemble correspondant à la détection de l'absence de $s_1(t)$, par

$$\forall f, \quad E_{\mathcal{T}_1}(f;k) = \frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1} E(t,f;k) - \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} E(t,f;k). \tag{5.28}$$

Ce nouveau profil moyen centré est donc indépendant de $\Lambda(f)$. Nous proposons donc maintenant, pour estimer l'ensemble des permutations $\mathcal{P}_f$, de rechercher à chaque fréquence f, parmi les N_s profils moyens centrés, le plus petit en valeur moyenne :

$$\forall f, \quad \widehat{\mathcal{P}}(f) = \arg\min_{\mathcal{P}(f)} \Big(\mathcal{P}(f)\,\mathbf{E}_{\mathcal{T}_1}(f;:)\Big)_1, \tag{5.29}$$

où $\mathbf{E}_{\mathcal{T}_1}(f;:) = [E_{\mathcal{T}_1}(f;1),\cdots,E_{\mathcal{T}_1}(f;N_s)]^T$ est le vecteur dont les composantes sont les N_s profils moyens centrés.

Finalement, nous reconstruisons la première source $\hat{S}_1(t,f)$ en appliquant à chaque fréquence f la matrice de permutation $\widehat{\mathcal{P}}(f)$ ainsi estimée :

$$\hat{S}_1(t,f) = \widehat{\mathcal{P}}(f)\,Y_1(t,f). \tag{5.30}$$

Le principe de cette régularisation des permutations est résumé dans l'algorithme 3.

Algorithme 3 Régularisation des permutations par un détecteur d'activité vocale.

/Estimation des moments de silence de $s_1(t)$/
Estimer $\mathcal{T}_1$ grâce au détecteur d'activité vocale visuel (chapitre 4)

/Estimation de l'ensemble des permutations/
Pour toutes les frequences f **faire**
  Calculer les N_s profils moyens centrés $E_{\mathcal{T}_1}(f;k)$, $k \in \{1,\cdots,N_s\}$ par (5.28)
  Estimer la permutation $\widehat{\mathcal{P}}(f)$ par (5.29)
Fin boucle

/Estimation de $s_1(t)$/
Estimer $\hat{S}_1(t,f)$ par (5.30)

Notons que ce principe permet de résoudre les permutations pour une source donnée si nous disposons d'un oracle associé indiquant les moments de silence de cette source comme par exemple le détecteur d'activité vocale visuel que nous avons introduit. Mais il peut rester des permutations non résolues sur les autres sources, ce qui n'a pas de conséquences sur l'extraction de $s_1(t)$. Extraire plus d'une source avec ce principe nécessitera d'avoir le(s) oracle(s) correspondant(s). Par exemple, si deux oracles associés à deux sources $s_1(t)$ et $s_2(t)$ fournissent $\mathcal{T}_1$ et $\mathcal{T}_2$ qui sont respectivement les ensembles pour lesquels les sources sont absentes, alors nous proposons d'adapter notre principe sans utiliser cependant leur intersection pour le calcul des profils moyens centrés (5.28). Ainsi, $\mathcal{T}_1$ est remplacé par $\mathcal{T}_1\backslash\mathcal{T}_2 = \mathcal{T}_1 - (\mathcal{T}_1 \cap \mathcal{T}_2)$ pour estimer les permutations de $s_1(t)$ et $\mathcal{T}_2\backslash\mathcal{T}_1 = \mathcal{T}_2 - (\mathcal{T}_2 \cap \mathcal{T}_1)$ pour

celles de $s_2(t)$, si ces ensembles sont non vides. Ainsi, par exemple, pour résoudre les permutations sur la première source, nous cherchons à chaque fréquence f la permutation qui vérifie

$$\forall f, \quad \widehat{\mathcal{P}}(f) = \arg\min_{\mathcal{P}(f)} \Big(\mathcal{P}(f)\, \mathbf{E}_{\mathcal{I}_1 \setminus \mathcal{I}_2}(f) \Big)_1,$$

où les composantes de $\mathbf{E}_{\mathcal{I}_1 \setminus \mathcal{I}_2}(f)$ sont définies par (5.28). Ne pas tenir compte des intersections des moments de silence doit permettre de garantir une plus grande différence pour les valeurs des N_s profils moyens centrés sur ces intervalles de façon à éviter au maximum les erreurs de détection des permutations. Plus généralement, si les mélanges comptent plus de deux sources, il est nécessaire d'avoir suffisamment de diversité dans l'évolution temporelle des profils. On est en droit de penser qu'une telle diversité existe dans une conversation spontanée entre plusieurs locuteurs.

5.4 Résultats expérimentaux

Dans ce paragraphe nous présentons tout d'abord les résultats expérimentaux de l'extraction de source par la cohérence audiovisuelle exprimée par le modèle audiovisuel, puis les résultats expérimentaux de l'extraction par la parcimonie grâce au détecteur de silence visuel.

5.4.1 Extraction par la cohérence audiovisuelle

Pour tester notre algorithme de régularisation des indéterminations exploitant le modèle audiovisuel, le corpus des logatomes présenté au paragraphe 3.4 sert pour la source d'intérêt $s_1(t)$. Les paramètres $\{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Gamma_i^A\}_{1 \leq i \leq N_{AV}}$ du modèle audiovisuel ont été appris par l'algorithme EM comme nous l'avons expliqué au paragraphe 3.3.4. La seconde source est choisie parmi le corpus des phrases présenté au paragraphe 3.4.

Nous rappelons que la fréquence d'échantillonnage des signaux acoustiques est de 16kHz et que les trames ont une longueur de 20ms soit 320 échantillons audio.

Résultats de l'estimation du facteur d'amplitude

Dans ce paragraphe, nous présentons les résultats de l'estimation du facteur d'amplitude par une détermination soit purement acoustique (5.18) soit audiovisuelle (5.23). Pour cela, dans cette expérience aucun mélange ni séparation n'ont été effectués. Le signal $S_1(t, f)$ (obtenu par TFCT de $s_1(t)$) a été multiplié arbitrairement par un facteur échelle $\Lambda(f)$ choisi aléatoirement à chaque fréquence f.

Pour quantifier les performances, nous définissons la distorsion pour la première source par

$$d_1(t) = \sqrt{\frac{1}{N_f} \sum_{i=1}^{N_f} \Big| \log |S_1(t, f_i)| - \log |\hat{S}_1(t, f_i)| \Big|^2}. \tag{5.31}$$

Une bonne estimation du facteur d'amplitude correspond à une distorsion qui tend vers zéro. Dans le cas spécifique de notre expérience, où il n'y a pas de permutation,

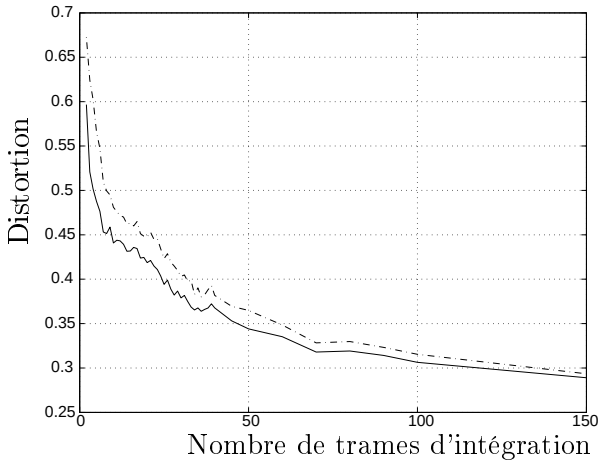


FIG. 5.4 – Distorsion (5.31) en fonction du nombre de trames d'intégration. Trait continu : estimation audiovisuelle (5.23) et trait discontinu : estimation audio (5.18).

la distorsion (5.31) ne dépend plus du temps t et peut être exprimée par

$$d_1 = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left|\log\big(\hat{\mathcal{L}}_{1,1}(f_i)\,\Lambda_{1,1}(f_i)\big)\right|^2}.$$

La figure 5.4 montre la valeur moyenne de la distorsion en fonction du nombre de trames d'intégration T pour les estimations audio (5.18) et audiovisuelle (5.23). Chaque simulation a été répétée 60 fois en sélectionnant pour chacune d'elle des logatomes différents. Heureusement, pour les deux estimations la distorsion décroît lorsque le nombre de trames d'intégration augmente. Ceci est dû au fait que la variance d'une section particulière de la parole peut être différente de la variance du modèle : plus la section de parole est longue, plus l'estimation est robuste. De plus, l'estimation audiovisuelle (5.23) est meilleure que celle audio (5.18), justifiant ainsi notre idée que la modalité visuelle peut être une aide efficace pour estimer la variance du modèle correspondant à une section particulière de la parole. Ainsi, pour une distorsion arbitraire, le nombre de trames d'intégration est plus petit pour une estimation audiovisuelle que pour une estimation audio.

Résultats de l'estimation des permutations

Pour estimer les performances de notre algorithme de régularisation des permutations, nous testons la détection de permutation pour des blocs de fréquences consécutives. Tout comme dans la section précédent, aucun mélange ni séparation n'ont été effectués. Nous avons simplement permuté artificiellement des blocs de fréquences consécutives entre les deux sources $S_1(t,f)$ et $S_2(t,f)$ obtenues par TFCT de $s_1(t)$ et $s_2(t)$ respectivement. Ensuite, nous appliquons notre algorithme de détection des permutations sur ces signaux artificiellement modifiés. Tout d'abord, nous avons permuté 1, 4, 8, 12 et 16 blocs de fréquences dont la largeur de bande est égale à 250Hz (soit regroupant 5 fréquences consécutives). Dans ce cas, la résolution maximale (correspondant au nombre de fréquences consécutives) que nous cherchons

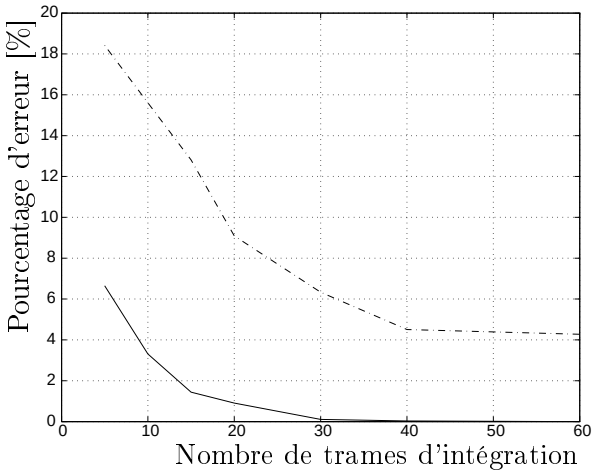
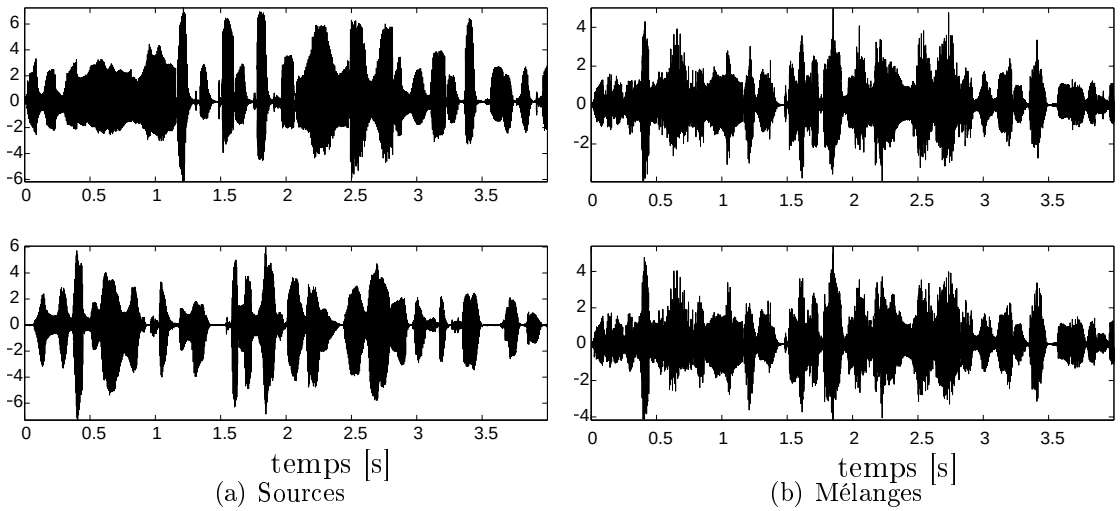


FIG. 5.5 – Pourcentage d'erreur de détection des permutations en fonction du nombre de trames d'intégration. Le trait continu correspond à la permutation des blocs d'une largeur de 250Hz et le trait discontinu à la permutation des blocs d'une largeur de 100Hz.

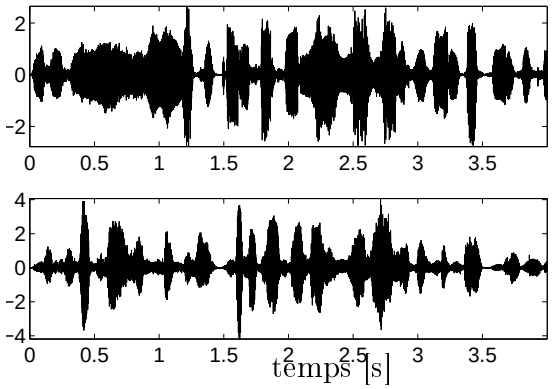
à détecter est au mieux de 250Hz. C'est-à-dire que nous ne testerons dans notre algorithme que des permutations d'au moins 5 fréquences consécutives. Ensuite, nous avons répété ce test mais en choisissant de permuter des blocs de fréquences consécutives d'une largeur égale à 100Hz (soit deux fréquences consécutives) pour 1, 10, 20, 30 et 40 blocs permutés dans le spectre. Dans ce cas la résolution maximale de notre algorithme pour détecter des permutations est de 100Hz.

Nous définissons le nombre d'erreur de détection des permutations comme la somme des permutations effectivement présentes mais non détectées par notre algorithme et des mauvaises permutations (c'est-à-dire des mauvaises décisions de notre algorithme).

Pour chaque condition d'expérimentation, les simulations ont été répétées 60 fois pour différents logatomes choisis aléatoirement mais connu de façon à avoir une référence. La figure 5.5 montre la valeur moyenne du pourcentage d'erreur de détection en fonction du nombre de trames d'intégration T pour les deux cas étudiés (blocs de largeur 100Hz ou 250Hz). Cette figure montre l'importance de l'intégration pour les critères (5.9) et (5.11). En effet, si le nombre de trames d'intégration diminue alors le taux d'erreurs de détection augmente mais le coût de calcul diminue puisqu'il y a moins de densités à calculer. Au contraire, si le nombre de trames d'intégration augmente alors le taux d'erreurs de détection diminue vers zéro mais le coût de calcul augmente. Notons que pour un nombre de trames d'intégration égal à 40, le pourcentage d'erreurs de détection est inférieur à 5% pour les deux conditions de test. De plus, la valeur moyenne des résultats pour les blocs de largeur 250Hz et toujours meilleure que celle obtenue pour les permutations des blocs de largeur 100Hz ce qui signifie qu'il est plus facile de détecter la permutation d'un bloc d'une grande largeur spectrale (*i.e.* regoupant un grand nombre de fréquences consécutives). Finalement, pour un pourcentage d'erreur de détection arbitraire, il est possible d'augmenter la résolution de l'algorithme mais au prix d'un plus grand nombre de trames d'intégration.



(a) Sources (b) Mélanges



(c) Sources estimées

FIG. 5.6 – Résultat de la séparation par le modèle audiovisuel.

Résultats pour l'extraction

Dans la suite de ce paragraphe, nous considérons le cas de deux sources (figure 5.6(a)) et de deux mélanges (figure 5.6(b)). Tous les filtres de mélanges sont des filtres artificiels à réponse impulsionnelle finie de 320 échantillons : ils modélisent de façon simplifiée la réponse impulsionnelle acoustique d'une pièce (figure 5.7). Même si les signaux sont tracés sur quatre secondes, nous n'avons utilisé que les 40 premières trames pour détecter les permutations. Cette valeur est cohérente avec celle obtenue au paragraphe précédent.

De façon à quantifier la qualité de l'estimation de la matrice de séparation, nous utilisons l'indice de performance [102], défini par

$$r_1(f) = \sum_{j=1}^{N_s} \frac{|C_{1,j}(f)|}{|C_{1,1}(f)|} - 1, \tag{5.32}$$

où $C(f) = G(f)\,H(f)$ est la matrice du système global. Pour une bonne séparation, cet index doit être proche de zéro (ou l'infini si une permutation a lieu).

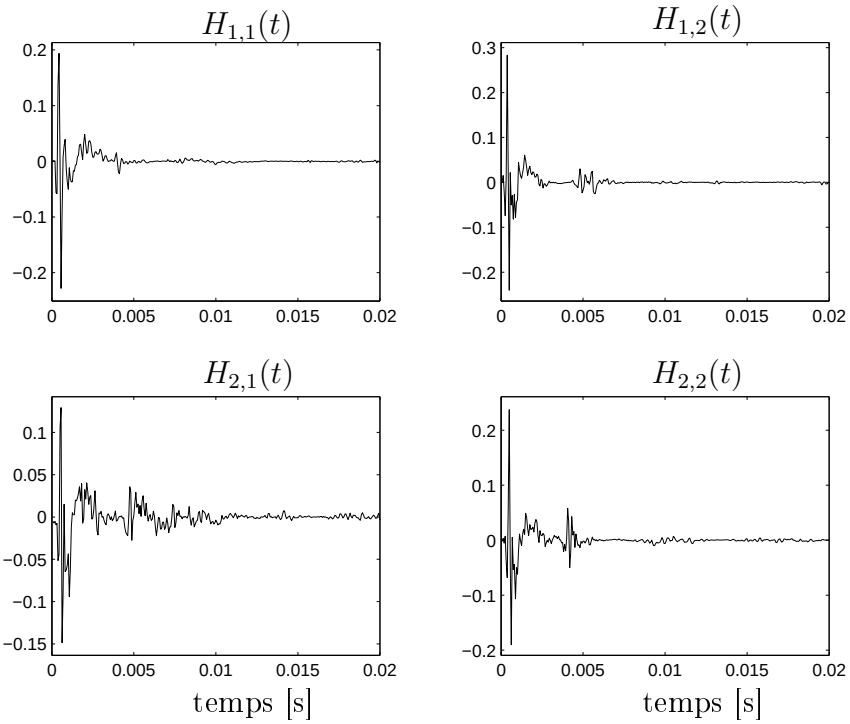


FIG. 5.7 – Réponses impulsionnelles des filtres de mélanges.

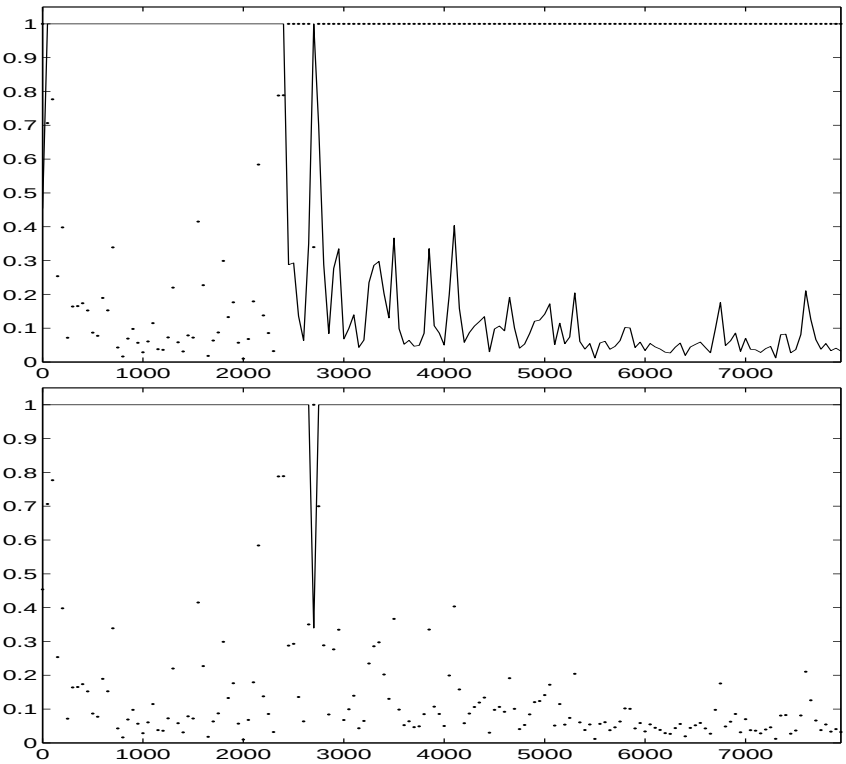


FIG. 5.8 – Indice de performance $r_1(f)$ (5.32) (pointillé) et son inverse (trait continue) tronqués à 1, avant (tracé du haut) et après (tracé du bas) notre algorithme de régularisation des permutations en fonction de la fréquence en Hz.

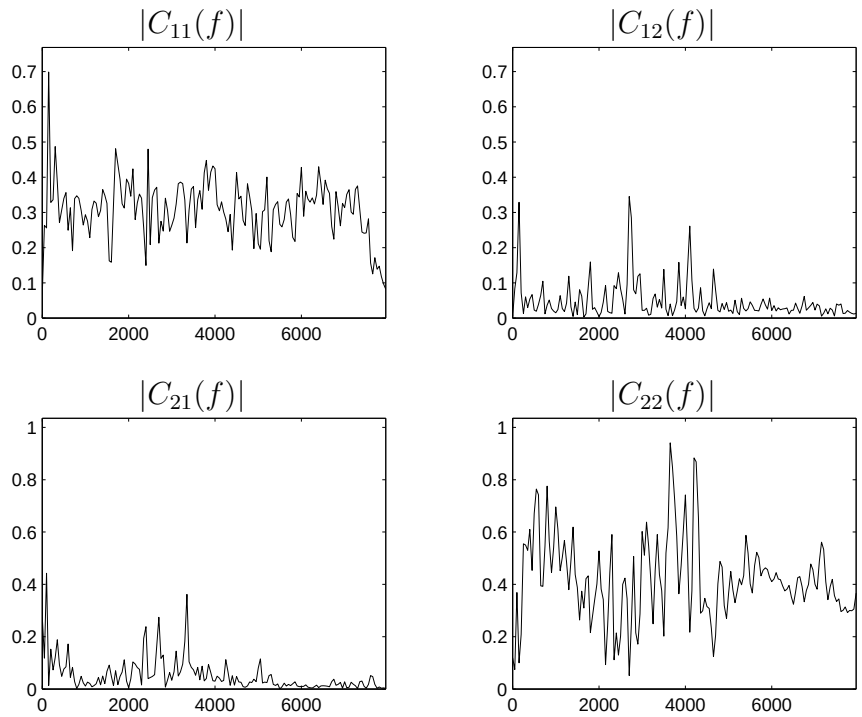


FIG. 5.9 – Réponse en fréquence du filtre global après l'estimation du facteur d'amplitude.

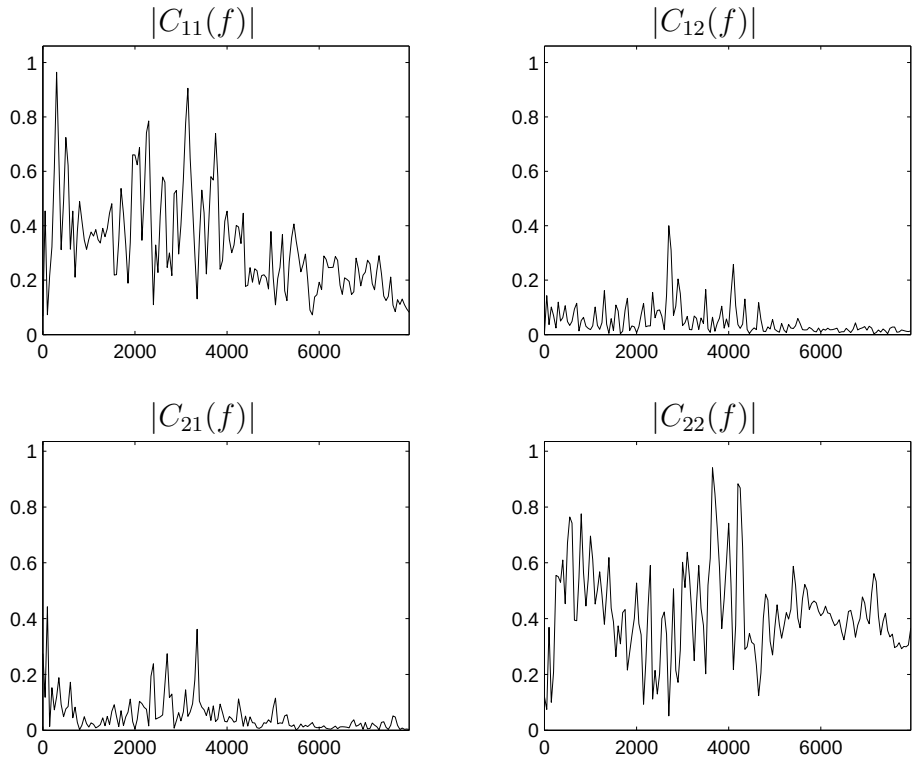


FIG. 5.10 – Réponse en fréquence du filtre global avant l'estimation du facteur d'amplitude.

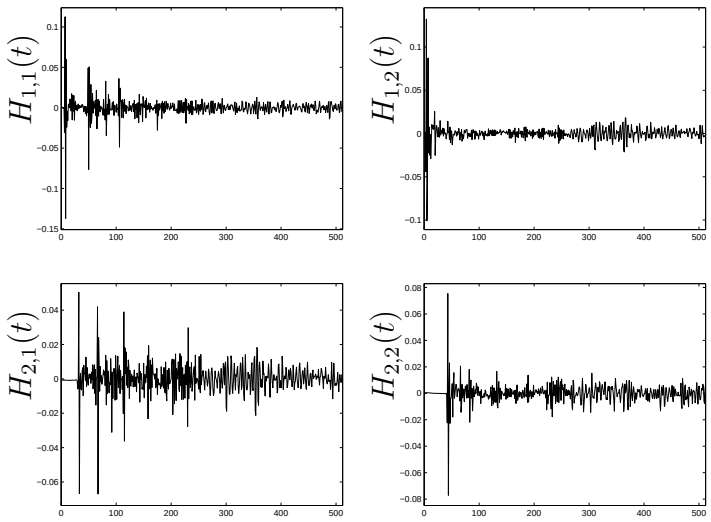


FIG. 5.11 – Réponse impulsionnelle des filtres de mélanges.

Sur la figure 5.8 sont tracés $\min(1, r_1(f))$ et $\min(1, 1/r_1(f))$ avant et après la détection des permutations en fonction de la fréquence f. On peut voir que le principe proposé a corrigé toutes les permutations exceptée une permutation : en effet, pour toutes les fréquences, l'indice de performance est plus petit que un, à l'exception d'une seule fréquence aux alentours de 2750Hz. Ceci est confirmé par le spectre des réponses en fréquence des filtres globaux $C(f)$ (*cf.* figure 5.9) : pour toutes les fréquences f (excepté pour l'erreur), $|C_{1,2}(f)|$ (resp. $|C_{2,1}(f)|$) est plus petit que $|C_{1,1}(f)|$ (resp. $|C_{2,2}(f)|$). Ceci signifie que la matrice de filtres globale $\big(G * H\big)(t)$ est proche d'une matrice diagonale de filtres.

Finalement, la figure 5.10 montre le spectre des réponses en fréquence du filtre global où nous avons seulement appliqué la détection des permutations : nous avons estimé l'ensemble des permutations par notre algorithme final présenté au paragraphe 5.2.3 et le filtre global est ici estimé par $\widehat{\mathcal{P}}(f)\, G(f)\, H(f)$ (*i.e* sans l'estimation du facteur d'amplitude). On peut voir que le spectre de $C_{1,1}(f)$ est plus proche d'une constante avec l'estimation du facteur d'amplitude (figure 5.9) que sans estimation du facteur d'amplitude (figure 5.10). Ceci entraîne une meilleure estimation de l'allure spectrale de la source d'intérêt. De plus, notons que $C_{2,1}(f)$ et $C_{2,2}(f)$ sont inchangés puisque nos critères ne portent que sur la première source.

5.4.2 Extraction par la parcimonie

Dans ce paragraphe, nous présentons les résultats obtenus par l'algorithme de régularisation des permutations par la parcimonie introduit au paragraphe 5.3. Nous considérons le cas de deux sources mélangées par des matrices de filtres 2×2. Ces filtres sont des filtres de réponse impulsionnelle finie de 512 coefficients avec trois échos principaux (figure 5.11). Ils sont extraits d'une bibliothèque de réponses impulsionnelles mesurées dans une grande pièce de 3.5m×7m×3m (on peut les trouver à l'adresse suivante `http ://sound.medi.mit.edu/ica-bench`). Le corpus utilisé

pour la source $s_1(t)$ à extraire est de la parole continue produite par un locuteur masculin enregistré en condition de dialogue spontané (issue du corpus "Grenoble"). La seconde source est de la parole continue produite par un autre locuteur (issue du corpus des phrases). Dans toutes nos expériences, la taille des TFCT est de 4096 échantillons.

Résultat de l'estimation des permutations

Nous donnons dans ce paragraphe les performances de notre principe pour régulariser les permutations par la parcimonie. Pour cela définissons tout d'abord la contribution d'un signal à un autre.

Définition 5.1 (Contribution d'un signal à un autre)
Soient $w(t)$, $y(t)$ et $z(t)$ trois signaux tels que $z(t) = f(y(t)) + w(t)$ où $f(\cdot)$ est une fonction et $w(t)$ ne dépend pas de $y(t)$. Nous appelons contribution de $y(t)$ à $z(t)$ le terme $f(y(t))$ et nous le notons $\big(y|z\big)(t)$: $\big(y|z\big)(t) = f(y(t))$.

Nous notons de plus P_y la puissance moyenne du signal $y(t)$

$$P_y = \frac{1}{T}\sum_{t=1}^{T} |y(t)|^2. \tag{5.33}$$

Ainsi, le rapport signal sur interférence (RSI) pour la première source est définie par

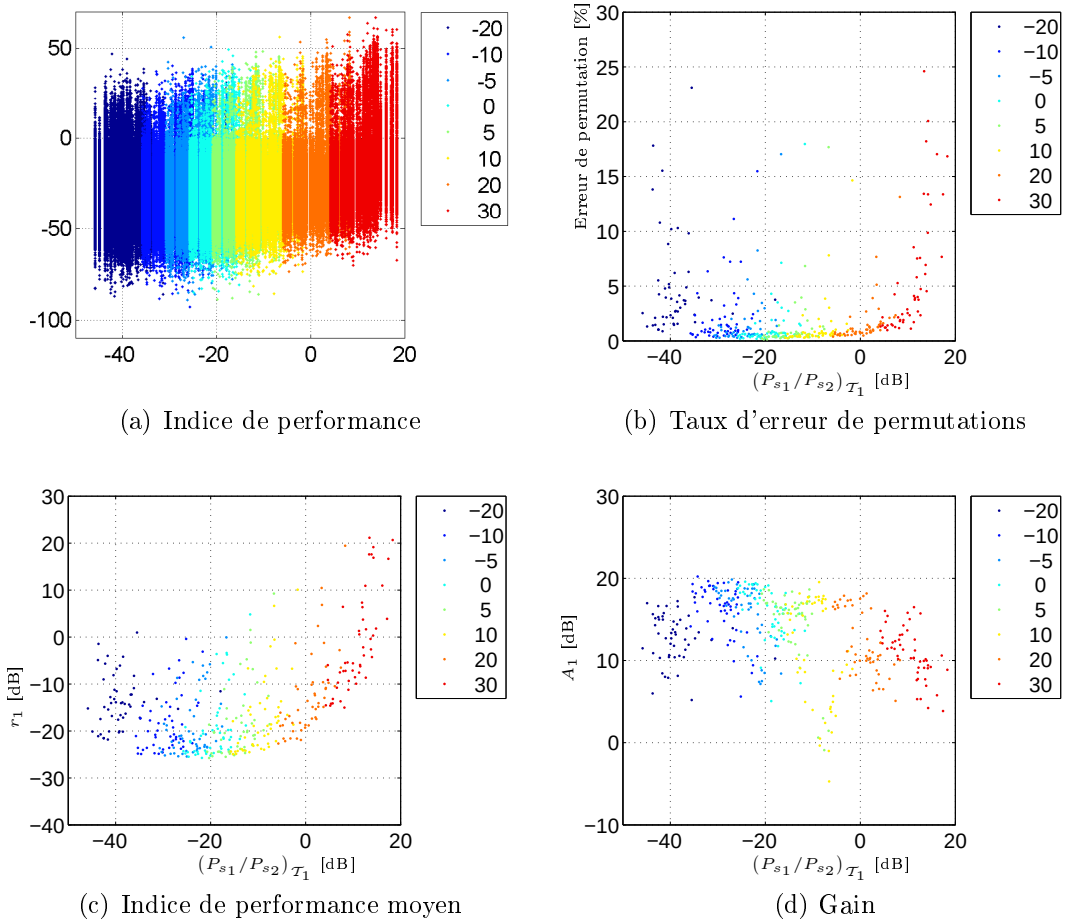
$$RSI_{(s_1|\hat{s}_1)} = \frac{P_{(s_1|\hat{s}_1)}}{\sum_{s_j \neq s_1} P_{(s_j|\hat{s}_1)}}. \tag{5.34}$$

Remarquons que $(s_j|\hat{s}_1)(t) = \sum_{i=1}^{N_o} G_{1,i}(t) * H_{i,j}(t) * s_j(t)$. Cet indice classique en séparation de source quantifie la qualité de l'estimation de la source $\hat{s}_1(t)$. Pour une bonne estimation de la source (*i.e.* $\forall j > 1,\ (s_j|\hat{s}_1)(t) \simeq 0$), cet indice doit tendre vers plus l'infini. Finalement, nous définissons le gain sur la première source grâce à la fonction de séparation par

$$A_1 = \frac{RSI_{(s_1|\hat{s}_1)}}{\max_l RSI_{(s_1|x_l)}} = \min_l \frac{RSI_{(s_1|\hat{s}_1)}}{RSI_{(s_1|x_l)}} \tag{5.35}$$

où $RSI_{(\cdot|\cdot)}$ est défini par (5.34) et $(s_j|x_l)(t) = H_{l,j}(t) * s_j(t)$. Ce gain permet de quantifier l'amélioration en terme de RSI avant et après la séparation. La référence avant la séparation est prise par rapport au mélange pour lequel la contribution de la source à extraire est la plus grande.

Dans cette série d'expériences, nous présentons les performances obtenues par le principe de régularisation des permutations par la parcimonie. Pour cela, à chaque expérience, la source $s_1(t)$ (resp. $s_2(t)$) correspond à 20 secondes de parole choisies aléatoirement parmi l'ensemble du corpus de "Grenoble" (resp. des phrases). Nous définissons le rapport signal sur bruit en entrée comme le rapport des puissances moyennes des deux sources pendant les moments indexés manuellement *non silence*. Ce rapport signal sur bruit en entrée varie de -20dB à 30dB et pour chacun de ces rapports, 50 simulations ont été effectuées. Les matrices de filtres de mélange sont



(a) Indice de performance

(b) Taux d'erreur de permutations

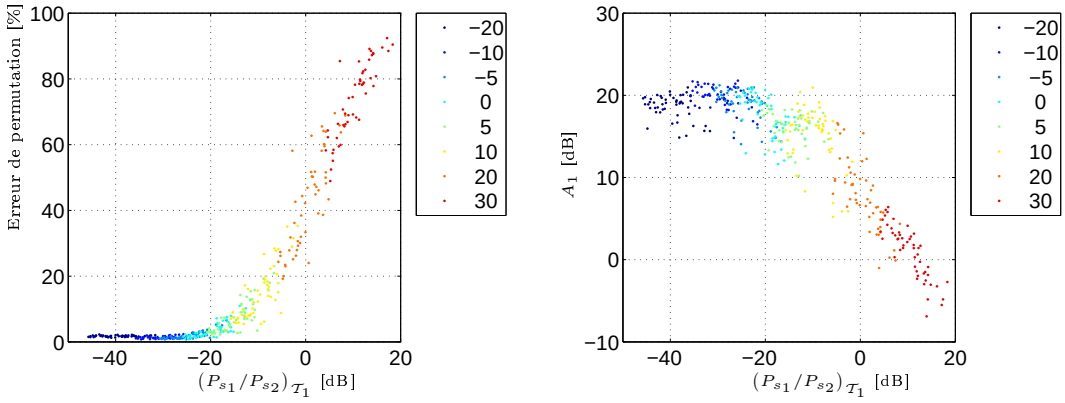
(c) Indice de performance moyen

(d) Gain

FIG. 5.12 – Performances de l'extraction par la parcimonie. Les légendes indiquent le rapport signal sur bruit en entrée. Figure 5.12(a) : indice de performance $r_1(f)$ en fonction du rapport des puissances des sources pendant les moments de silence $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ (exprimé en dB). Figure 5.12(b) : taux d'erreur de détection des permutations (en %) en fonction de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$. Figure 5.12(c) indice de performance moyen $r_1 = 1/N_f \sum_{i=1}^{N_f} r_1(f_i)$ en fonction de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$. Figure 5.12(d) : gain A_1 (5.35) en fonction de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$.

gardées constantes et les réponses impulsionnelles de ceux-ci sont données à la figure 5.11. La matrice des filtres de séparation $G(f)$ est estimée à chaque fréquence f par diagonalisation conjointe des matrices de densité spectrale à court terme des observations. Nous appliquons ensuite le principe de régularisation des permutations résumé dans l'algorithme 3, où la détermination de l'ensemble des moments de silence $\mathcal{T}_1$ est ici remplacée par l'indexation manuelle des silences, que nous avons fait au préalable. Les performances, en terme d'indice de performance $r_1(f)$, d'indice de performance moyen $r_1 = 1/N_f \sum_{i=1}^{N_f} r_1(f_i)$, de gain A_1 et de taux d'erreur de détection des permutations, sont données à la figure 5.12. Ces résultats sont tracés en fonction du rapport des puissances moyennes des deux sources, P_{s_1} et P_{s_2} calculées pendant les instants indexés *silence* $\mathcal{T}_1$, rapport que l'on note $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$. A la figure 5.12(a), pour chaque valeur de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$, est représenté l'indice de performance $r_1(f)$ pour toutes les fréquences, la figure 5.12(c) reportant quant à elle l'indice de performance moyen r_1. Bien que pour certaines fréquences l'indice $r_1(f)$ soit supérieur à 1 (ou 0 si celui-ci est exprimé en dB), l'estimation de la matrice de séparation est correcte comme le montre d'une part le gain A_1 qui est supérieur à 1 dans la majeure partie des cas et d'autre part l'indice de performance moyen r_1 qui reste inférieur à 1. Quand le rapport $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ devient important (très supérieur à 1), l'indice de performance moyen r_1 prend une valeur de l'ordre de 1 (ou supérieure à 1) ce qui devrait impliquer une mauvaise estimation de la source à extraire. Cependant, comme le montre la figure 5.12(d), le gain A_1 est d'environ 10dB. Pour donner une explication, regardons la figure 5.12(b) qui reporte le pourcentage d'erreur de détection des permutations déterminées par un indice de performance $r_1(f)$ supérieur à 1. On constate que pour environ 10% des fréquences, $r_1(f)$ est plus grand que 1 et que dans le même temps pour la majorité des fréquences $r_1(f)$ est très inférieur à 1 comme on peut le voir à la figure 5.12(a). Finalement, notre algorithme de régularisation des permutations fonctionne correctement comme on peut le voir à la figure 5.12(b) puique, dans la majeure partie des cas, il y a moins de 5% des fréquences pour lesquelles il reste une permutation résiduelle. De plus, ces permutations résiduelles n'ont que peu d'influence sur l'estimation de la source à extraire comme le montre le gain compris entre 10 et 20dB.

Il est intéressant de remarquer que notre algorithme est fondé sur la recherche de la puissance moyenne minimale pendant les moments de silence de la source à extraire. Cependant, même lorsque $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ (le rapport des puissances moyennes entre les sources $s_1(t)$ et $s_2(t)$ pendant les moments de silence $\mathcal{T}_1$) est supérieur à 1, notre algorithme associe correctement les composantes spectrales correspondant à la source à extraire comme le montre la figure 5.12(b). Ceci s'explique par l'emploi des profils moyens centrés $E_{\mathcal{T}_1}(f;k)$ (5.28) et non pas directement des profils moyens $E(t,f;k)$ calculés pour $t \in \mathcal{T}_1$. En effet, en notant $P_{s_1}(\mathcal{T}_1)$ la puissance moyenne calculée pendant $\mathcal{T}_1$, $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1} > 1$ implique que $P_{s_1}(\mathcal{T}_1) > P_{s_2}(\mathcal{T}_1)$. Or cette situation ne se produit que lorsque le RSB en entrée est supérieur à 20dB (*cf.* figure 5.12(a)), c'est-à-dire que $P_{s_1}(\mathcal{T}) \gg P_{s_2}(\mathcal{T})$, où $\mathcal{T}$ est l'ensemble des indices temporels. Faisons maintenant un ordre de grandeur qualitatif en supposant que la répartition de la puissance est uniforme pour toutes les fréquences et n'évolue pas au court du temps.



(a) Taux d'erreur de permutations (b) Gain

FIG. 5.13 – Performances de l'extraction par la parcimonie qui exploite les valeurs moyennes des profils non centrés. Les légendes indiquent le rapport signal sur bruit en entrée. Figure 5.13(a) : taux d'erreur de détection des permutations (en %) en fonction de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$. Figure 5.13(b) : gain A_1 en fonction de $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$.

On peut alors réécrire $P_{s_1}(\mathcal{T}_1) > P_{s_2}(\mathcal{T}_1)$ comme

$$10\log P_{s_1}(\mathcal{T}_1) \# 10\log P_{s_2}(\mathcal{T}_1) + qq\, dB,$$

où $\#$ signifie "de l'ordre de" et qq signifie "quelques" (c'est-à-dire que qq prend une valeur comprise entre environ 1 et 3). L'équation précédente se lit alors "la puissance P_{s_1} exprimée en décibel est de l'ordre de P_{s_2} en dB plus quelques dB". Or $E(t,f;1) \# \ln P_{s_1}(\mathcal{T}_1)$ et $E(t,f;2) \# \ln P_{s_2}(\mathcal{T}_1)$ si l'on suppose qu'il n'y a pas de permutations. D'après l'équation précédente on obtient que $E(t,f;1) > E(t,f;2)$ ce qui implique que l'algorithme, qui exploiterait les profils moyens non centrés, détecte une permutation à cette fréquence. D'autre part, on a vu que $P_{s_1}(\mathcal{T}_1) > P_{s_2}(\mathcal{T}_1)$ implique que $P_{s_1}(\mathcal{T}) \gg P_{s_2}(\mathcal{T})$ ce que l'on peut réécrire comme

$$10\log P_{s_1}(\mathcal{T}) \# 10\log P_{s_2}(\mathcal{T}) + QQ\, dB,$$

où QQ signifie un "gros quelques" (c'est-à-dire une valeur comprise entre 6 et 9). Ainsi, en supposant qu'il n'y ait pas de permutation à la fréquence considérée, à une constante multiplicative près on a

$$\begin{aligned} E_{\mathcal{T}_1}(f;1) &\# 10\log P_{s_1}(\mathcal{T}_1) - 10\log P_{s_1}(\mathcal{T}) \\ &\# \left(10\log P_{s_2}(\mathcal{T}_1) + qq\, dB\right) - \left(10\log P_{s_2}(\mathcal{T}) + QQ\, dB\right) \\ &\# 10\log P_{s_2}(\mathcal{T}_1) - 10\log P_{s_2}(\mathcal{T}) - qq\, dB \\ &\# E_{\mathcal{T}_1}(f;2) - qq\, dB, \end{aligned}$$

et finalement $E_{\mathcal{T}_1}(f;1) < E_{\mathcal{T}_1}(f;2)$. Donc l'algorithme, qui exploite les profils moyens centrés, ne détecte pas de permutation à cette fréquence d'où un nombre d'erreur de détection des permutations plus faible avec le principe que nous proposons que si nous avions utilisé les valeurs moyennes des profils pendant $\mathcal{T}_1$ sans les centrer comme

illustré à la figure 5.13. Pour conclure, le fait d'employer les profils moyens centrés permet non seulement de ne plus dépendre du facteur d'amplitude inconnu, mais cela engendre de plus un algorithme plus robuste puisqu'il produit moins d'erreurs de détection de permutation.

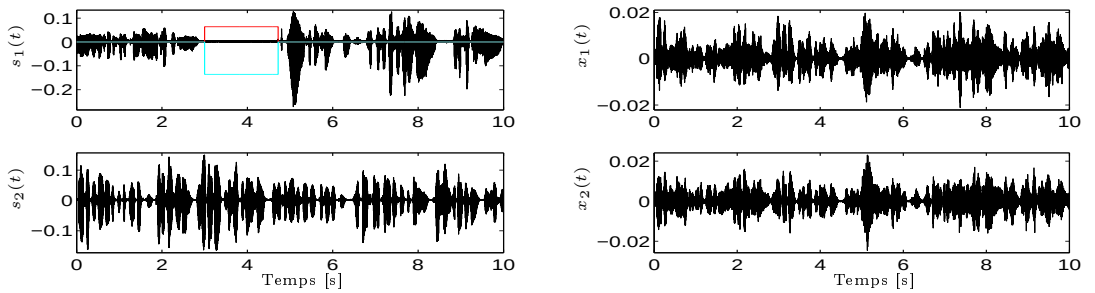
Résultat pour l'extraction

La figure 5.14 présente un exemple de séparation. Les Figures 5.14(a) et 5.14(b) montrent les deux sources et les mélanges. Dans ces expériences, dix secondes de signal sont utilisées pour estimer des filtres de séparation de 4096 coefficients (ce qui est donc la taille de toutes les TFCT). Les Figures 5.14(e) et 5.14(g) montrent les sources estimées dans différentes conditions (voir ci-dessous). Les indices $r_1(f)$ correspondants, tronqués à 1, sont représentés sur les Figures 5.14(f) et 5.14(h). Sur la figure 5.14(a) le trait rouge représente une indexation manuelle des silences et le trait cyan représente la détection automatique obtenue avec le détecteur de silence visuel du paragraphe 4.3.1.

Dans la première expérience (Figures 5.14(e) et 5.14(f)), les sources sont estimées par l'algorithme de diagonalisation conjointe sans régularisation des permutations. On peut voir que plusieurs blocs de fréquences consécutives sont permutés, ainsi que plusieurs fréquences isolées ce qui se traduit par un indice de performance $r_1(t)$ supérieur à 1 (*cf.* figure 5.14(f)). Par conséquent, les signaux séparés contiennent des composantes basses/hautes fréquences permutées entre les deux sources (*cf.* figure 5.14(e)). Dans la seconde expérience (*cf.* figure 5.14(g) et 5.14(h)), les sources sont estimées en utilisant le détecteur de silence associé au locuteur 1 pour détecter les silences de $s_1(t)$ et ainsi régulariser les permutations en utilisant la technique du paragraphe 5.3. A partir des résultats du chapitre 4, on a choisi $\alpha_l = 0.82^l$ pour les paramètres d'intégration du détecteur de silence visuel et le nombre minimal de trames de silence consécutives, que l'on s'autorise à détecter, L est de 20 (*i.e.* la longueur minimum d'un silence est de 400ms). On peut voir que les permutations principales sont corrigées, ce qui permet une bonne estimation des sources. Il reste quelques permutations isolées mais une investigation plus profonde révèle qu'elles correspondent à des régions des spectres avec une très faible énergie pour les deux sources : elles ont donc une influence très faible sur la qualité de la séparation, comme on peut le voir à la figure 5.14(h) : $C_{1,1}(t)$ est largement supérieur à $C_{1,2}(t)$. Les Figures 5.14(d) et 5.14(c) montrent les profils centrés des deux sources estimées avant et après la correction des permutations par le DAV-V. On voit que les blocs permutés sont bien détectés par les profils centrés calculés à partir de la détection de silence : après la régularisation de permutations, on a $E_{\mathcal{T}_1}(f;1) \leq E_{\mathcal{T}_1}(f;2)$ ce qui conduit à une bonne estimation des sources. Ces observations sont confirmées par l'écoute des signaux.

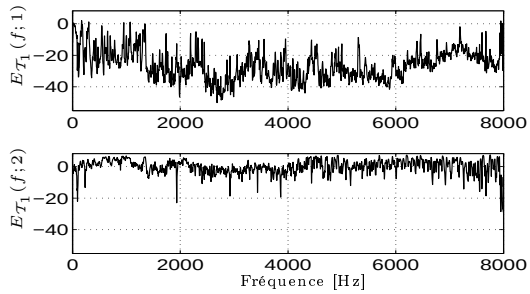
5.5 Conclusion

Dans ce chapitre, nous venons de proposer deux nouveaux algorithmes pour extraire le signal d'un locuteur particulier de mélanges convolutifs. La bimodalité de la parole est exploitée pour résoudre le problème des indéterminations qui découlent

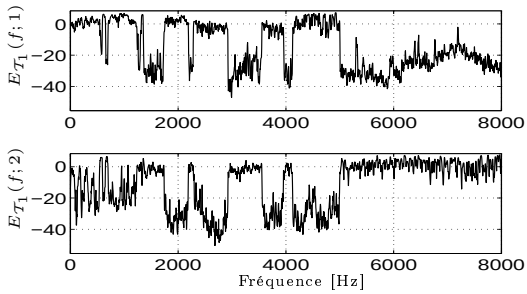


(a) Sources

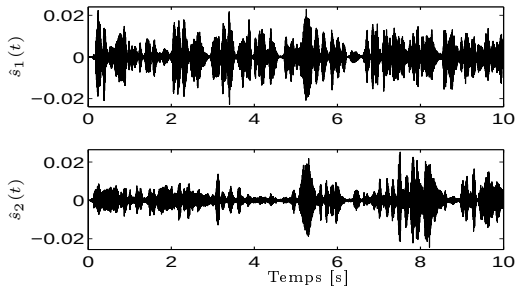
(b) Mélanges



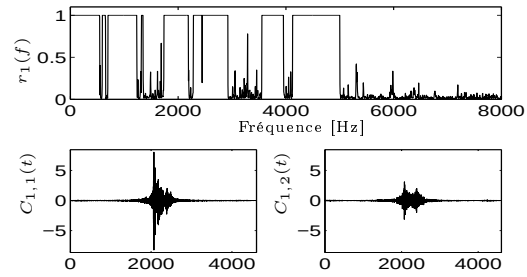
(c) Après régularisation des permutations



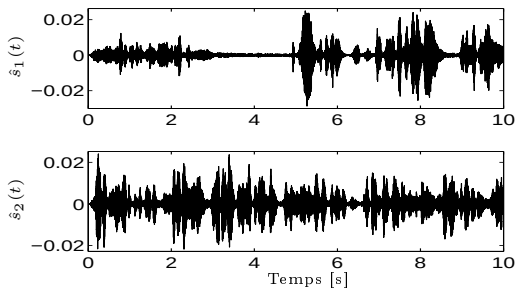
(d) Avant régularisation des permutations



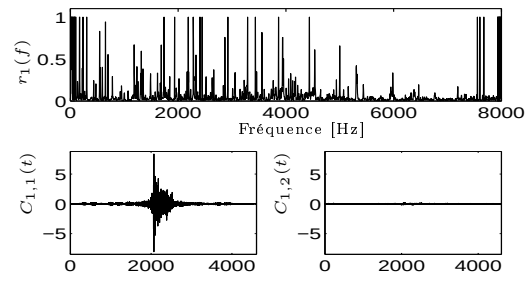
(e) Avant régularisation des permutations



(f) Avant régularisation des permutations



(g) Après régularisation des permutations



(h) Après régularisation des permutations

FIG. 5.14 – Résultat de l'extraction par la parcimonie. Sont tracés : les sources Fig.5.14(a) (le trait rouge est l'indexation manuelle du silence, le trait cyan le résultat du détecteur de silence visuel), les mélanges Fig. 5.14(b), les profils avant (Fig 5.14(d)) et après (Fig. 5.14(c)) la régularisation des permutations, les sources estimées avant (Fig. 5.14(e)) et après (Fig. 5.14(g)) la régularisation des permutations, les indices de performance ainsi que la première ligne du filtre global avant (Fig. 5.14(f)) et après (Fig. 5.14(h))la régularisation des permutations.

de la séparation, purement acoustique, fondée sur l'indépendance des sources. La première technique cherche à reconstruire un signal le plus cohérent avec l'observation vidéo du locuteur en exploitant le modèle audiovisuel du chapitre 3. Bien que performante, cette méthode n'en demeure pas moins gourmande en temps de calcul et souffre de l'apprentissage *a priori* du modèle audiovisuel qui dépend du locuteur. La seconde méthode repose quant-à-elle sur la parcimonie du signal de parole et sur la possibilité de détecter ces moments de silence grâce au détecteur d'activité vocal visuel du chapitre 4. Ainsi, les périodes de silence du locuteur particulier, rendant possible l'identification des permutations pour ce signal, permettent d'extraire ce locuteur quand il parle. Les expériences montrent que cette technique moins coûteuse en temps de calcul que la précédente est efficace même dans des conditions réalistes avec des filtres de mélanges contenant de nombreux échos.

Chapitre 6

Extraction directe par la parcimonie

Dans ce chapitre, nous présentons une nouvelle méthode fondée sur des considérations géométriques pour extraire une source de parole en exploitant la parcimonie de celle-ci, c'est-a-dire le fait que la parole spontanée comporte des moments de silence [110]. En effet, comme nous l'avons vu au chapitre précédent, un des inconvénients majeurs des systèmes de séparation dans le domaine fréquentiel fondés sur l'indépendance est de devoir régulariser les permutations rencontrées à chaque fréquence f. Alternativement, d'autres méthodes exploitent la parcimonie des sources. Par exemple, Abrard et Deville [1] proposent une solution dans le cas de mélanges instantanés. Ils exploitent le fait que dans le plan temps-fréquence, s'il existe des zones où une seule source est présente, alors il est possible de déterminer la matrice de mélange. Cependant, cette méthode a une restriction : elle requiert à la fois l'existence et la détection de ces zones temps-fréquence où une seule source est présente, ce qui est une hypothèse très forte. Récemment, Babaie-Zadeh *et al.* [9, 8] ont proposé une méthode géométrique dans le cas de mélanges instantanés de sources parcimonieuses. Leur méthode est fondée sur l'identification de la direction principale de la source présente dans les mélanges. Notre méthode est aussi fondée sur des caractéristiques géométriques, mais elle diffère de la leur puisque i) notre approche suppose que seule la source que l'on cherche à extraire doit être parcimonieuse, ii) l'indexation des sections, où la source à extraire est absente, est faite grâce au détecteur d'activité vocale visuel du chapitre 4, iii) nous traitons le cas des mélanges convolutifs.

Nous expliquons dans un premier temps le principe d'extraction dans le cas simple des mélanges complexes instantanés avant de l'étendre au cas des mélanges convolutifs. En effet, comme nous l'avons vu au chapitre 2, la résolution d'un problème de séparation de mélange convolutif peut se ramener à celle de plusieurs problèmes de séparation de mélange instantané de grandeurs spectrales complexes.

6.1 Cas des mélanges instantanés complexes

Pour expliquer le principe d'extraction basé sur la parcimonie de la parole que nous proposons, nous considérons d'abord le cas des mélanges linéaires instantanés de grandeurs complexes. Soient N_s sources indépendantes centrées à valeurs complexes $\mathbf{s}(t) \in \mathbb{C}^{N_s}$ et autant d'observations complexes $\mathbf{x}(t) \in \mathbb{C}^{N_o}$ ($N_o = N_s$)

obtenues par une matrice de mélange complexe $H \in \mathbb{C}^{N_s \times N_s}$:

$$\mathbf{x}(t) = H\,\mathbf{s}(t) \tag{6.1}$$

où $\mathbf{s}(t) = [s_1(t), \cdots, s_{N_s}(t)]^T$ et $\mathbf{x}(t) = [x_1(t), \cdots, x_{N_s}(t)]^T$. On suppose de plus que H est inversible. Ainsi, pour extraire toutes les sources, nous devons estimer une matrice de séparation $G \in \mathbb{C}^{N_s \times N_s}$ séparante :

$$\mathbf{y}(t) = G\,\mathbf{x}(t) \tag{6.2}$$

où $\mathbf{y}(t) = [y_1(t), \cdots, y_{N_s}(t)]^T$ est le vecteur des sources estimées. La matrice G étant séparante, celle-ci est égale à l'inverse de la matrice de mélange H à une permutation Π et un facteur d'échelle Λ près

$$G = \Pi\,\Lambda\,H^{-1}.$$

Pour extraire uniquement la première source $s_1(t)$, seule une ligne de G est nécessaire, nous choisissons arbitrairement la première notée $G_{1,:} = [G_{1,1}, \cdots, G_{1,N_s}]$:

$$y_1(t) = G_{1,:}\,\mathbf{x}(t) = G_{1,:}\,H\,\mathbf{s}(t) = C_{1,1}\,s_1(t), \tag{6.3}$$

où $C = G \times H$ est la matrice du système global. Dans la suite, nous proposons une nouvelle méthode afin d'estimer $G_{1,:}$. De plus, nous allons encore plus loin en régularisant le gain de sorte que la source $s_1(t)$ soit estimée au coefficient $H_{1,1}$ près, le coefficient de mélange, au lieu du coefficient arbitraire $C_{1,1}$. Ceci correspond à la situation où l'estimation de la source $s_1(t)$ serait égale à l'observation $x_1(t)$ alors que toutes les autres sources seraient absentes (*i.e.* $\forall i \neq 1,\ s_i(t) = 0$) : en d'autres termes, $s_1(t)$ est estimée au coefficient "canal+capteur" de l'observation considérée.

Pour estimer $G_{1,:}$, nous proposons une nouvelle méthode fondée sur des considérations géométriques. Soit $\mathcal{E}_s$ l'espace engendré par les sources $\mathbf{s}(t)$, ce que nous noterons

$$\mathcal{E}_s = \mathrm{Vec}\big\{\mathbf{s}(t)\big\}. \tag{6.4}$$

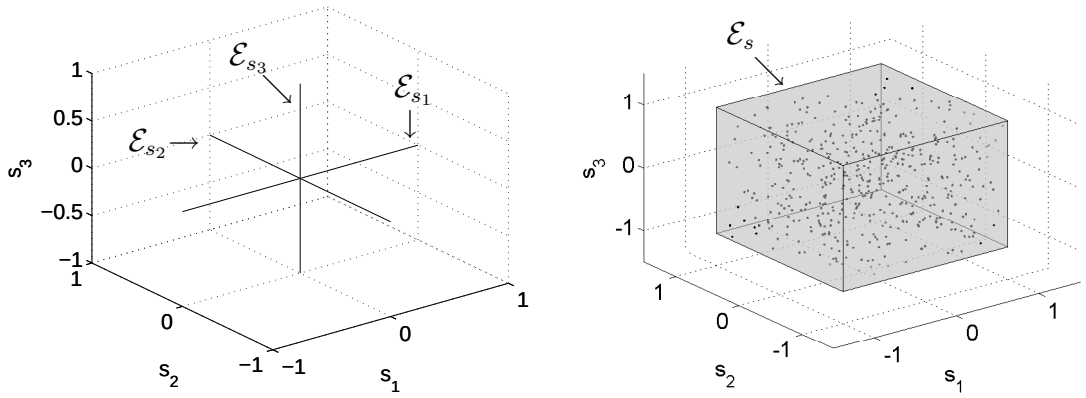
Or nous pouvons écrire le vecteur des sources sous la forme

$$\mathbf{s}(t) = \sum_{i=1}^{N_s} s_i(t)\,\mathbf{e}_i$$

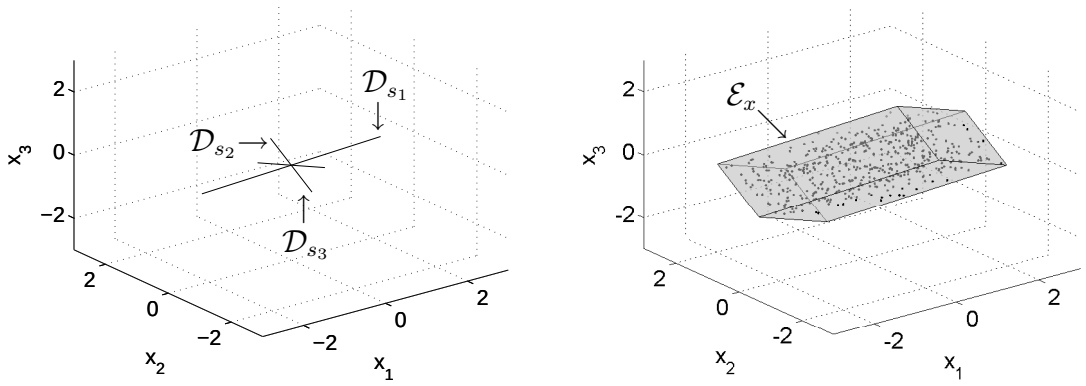
où $\mathbf{e}_i$ est le vecteur dont tous les coefficients sont nuls sauf le $i^{\text{ème}}$ qui vaut 1. Ainsi, les N_s sources étant indépendantes, $\mathcal{E}_s$ est la somme directe orthogonale des N_s espaces $\mathcal{E}_{s_i} = \mathrm{Vec}\big\{s_i(t)\,\mathbf{e}_i\big\}$, pour $1 \leq i \leq N_s$, engendrés par chacune des $s_i(t)$ sources :

$$\mathcal{E}_s = \bigoplus_{1 \leq i \leq N_s}^{\perp} \mathcal{E}_{s_i}. \tag{6.5}$$

Donc $\mathcal{E}_s$ est de dimension N_s. Notons que pour tout $1 \leq i \leq N_s$, $\mathcal{E}_{s_i}$ est défini sur un sous-ensemble de $\mathbb{C}$ dépendant de la distribution de la source correspondante. Par exemple, dans le cas de trois sources réelles mutuellement indépendantes et distribuées uniformément dans $[-1, 1]$, les $\mathcal{E}_{s_i}$ sont des espaces de dimension un



(a) Espace des sources (b) Espace des sources



(c) Espace des observations (d) Espace des observations

FIG. 6.1 – Illustration des espaces engendrés par trois sources indépendantes uniformément distribuées dans $[-1, 1]$.

défini sur les segments $[-1, 1]$ (*cf.* figure 6.1(a)). Donc l'espace $\mathcal{E}_s$ est de dimension trois et le volume occupé est un cube (*cf.* figure 6.1(b)).

Soit maintenant $\mathcal{E}_x$ l'espace engendré par les N_s mélanges $\mathbf{x}(t)$, il est obtenu à partir de $\mathcal{E}_s$ par la transformation linéaire définie par la matrice de mélange H

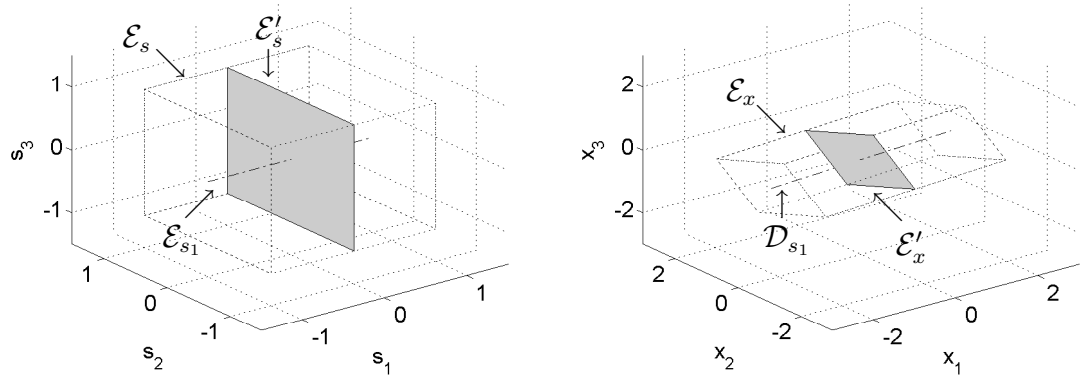
$$\mathcal{E}_x = \mathrm{Vec}\big\{\mathbf{x}(t)\big\} = \mathrm{Vec}\big\{H\,\mathbf{s}(t)\big\} = H\,\mathcal{E}_s. \tag{6.6}$$

puisque nous pouvons écrire que

$$\mathbf{x}(t) = \sum_{i=1}^{N_s} s_i(t)\,H_{:,i}. \tag{6.7}$$

Nous en déduisons que l'espace $\mathcal{D}_{s_i}$ engendré par la $i^{\text{ème}}$ source $s_i(t)$ dans l'espace des mélanges $\mathcal{E}_x$ est une droite dont la direction correspond à la $i^{\text{ème}}$ colonne de H

$$\mathcal{D}_{s_i} = \mathrm{Vect}\big\{s_i(t)\,H_{:,i}\big\}.$$



(a) Espace des sources (b) Espace des observations

FIG. 6.2 – Illustration des espaces engendrés par trois sources $\mathbf{s}(t)$ indépendantes uniformément distribuées dans $[-1, 1]$ lorsque s_1 s'annule.

L'espace $\mathcal{E}_x$ est alors de dimension N_s puisque somme directe des $\mathcal{D}_{s_i}$:

$$\mathcal{E}_x = \bigoplus_{1 \leq i \leq N_s} \mathcal{D}_{s_i}. \tag{6.8}$$

Dans notre exemple des trois sources réelles indépendantes et uniformément distribuées dans $[-1, 1]$, $\mathcal{E}_x$ est alors un espace de dimension trois et le volume occupé est un parallélépipède (*cf.* figure 6.1(d)).

Notons que l'écriture de l'équation des mélanges (6.1) sous la forme (6.7) ne permet de déterminer à une constante près la colonne $H_{:,1}$ de la matrice de mélange H correspondant à cette source, comme l'ont proposé Abrard et Deville [1], que si seule la source $s_1(t)$ est présente dans les mélanges $\mathbf{x}(t)$. Nous allons quant à nous adopter une démarche duale en cherchant à estimer directement la ligne $G_{1,:}$ de la matrice de séparation G permettant d'extraire la première source. Nous avons

$$y_1(t) = G_{1,:}\,\mathbf{x}(t) = \sum_{i=1}^{N_s} s_i(t)\, G_{1,:}\, H_{:,i}$$

que nous pouvons réécrire sous la forme

$$y_1(t) = C_{1,1}\, s_1(t) + \sum_{i=2}^{N_s} s_i(t)\, G_{1,:}\, H_{:,i}.$$

Ainsi, extraire la première source $s_1(t)$ implique que le vecteur $G_{1,:}$ soit orthogonal à tous les vecteurs $H_{:,i}$, pour $2 \leq i \leq N_s$:

$$y_1(t) \propto s_1(t) \Longleftrightarrow \forall\ i \geq 2,\ G_{1,:} H_{:,i} = 0. \tag{6.9}$$

Pour estimer $G_{1,:}$, supposons maintenant qu'un oracle, obtenu grâce au détecteur de silence visuel du chapitre 4, nous donne $\mathcal{T}_1$, un ensemble d'indices temporels où

$s_1(t)$ s'annule. D'après (6.5) l'espace $\mathcal{E}'_s$ engendré par les sources $\mathbf{s}(t)$ pour $t \in \mathcal{T}_1$ est un hyper-plan (*i.e.* un espace de dimension $N_s - 1$) de $\mathcal{E}_s$ dont le supplémentaire orthogonal est $\mathcal{E}_{s_1}$

$$\mathcal{E}_s = \mathcal{E}'_s \overset{\perp}{\oplus} \mathcal{E}_{s_1}.$$

De même, d'après (6.8), l'espace $\mathcal{E}'_x$ engendré, dans $\mathcal{E}_x$, par les sources $\mathbf{s}(t)$ pour $t \in \mathcal{T}_1$ est un hyper-plan dont $\mathcal{D}_{s_1}$ est un supplémentaire (non nécessairement orthogonal)

$$\mathcal{E}_x = \mathcal{E}'_x \oplus \mathcal{D}_{s_1}.$$

Notons que $\mathcal{E}'_s$ et $\mathcal{E}'_x$ sont les espaces engendrés par les $N_s - 1$ sources $s_i(t)$, avec $2 \leq i \leq N_s$, dans respectivement $\mathcal{E}_s$ et $\mathcal{E}_x$. Pour assurer que $G_{1,:}$ est orthogonal à tout $H_{:,i}$ pour $i \geq 2$ et ainsi extraire la première source $s_1(t)$, il suffit de projeter les observations $\mathbf{x}(t)$ parallèlement à $\mathcal{E}'_x$ sur un de ses supplémentaires, mais pas nécessairement sur $\mathcal{D}_{s_1}$.

Pour trouver un tel supplémentaire, nous proposons d'avoir recours à l'analyse en composante principale. En effet, on propose d'effectuer une décomposition en valeur propre de la matrice de covariance des observations $\mathbf{x}(t)$ pour $t \in \mathcal{T}_1$ (l'ensemble des indices temporels durant lequel $s_1(t)$ s'annule) : $C_{\mathbf{xx}}(\mathcal{T}_1) = \mathrm{E}_{t\in\mathcal{T}_1}[\mathbf{x}(t)\mathbf{x}^+(t)]$, où $^+$ est le transposé conjugué. Cette décomposition fournit N_s vecteurs propres orthogonaux $\mathbf{g}_i$, puisque $C_{\mathbf{xx}}(\mathcal{T}_1)$ est hermitienne, associés à N_s valeurs propres κ_i (qui sont triées par ordre décroissant : $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_{N_s}$). Les $N_s - 1$ vecteurs propres $\{\mathbf{g}_1, \cdots, \mathbf{g}_{N_s-1}\}$, associés aux $N_s - 1$ plus grandes valeurs propres $\{\kappa_1, \cdots, \kappa_{N_s-1}\}$, ont des directions qui appartiennent à $\mathcal{E}'_x$ tandis que le vecteur propre $\mathbf{g}_{N_s}$, associé à la plus petite valeur propre κ_{N_s}, est orthogonal à $\mathcal{E}'_x$. Ce vecteur propre $\mathbf{g}_{N_s}$ définit donc la direction du supplémentaire orthogonal $\mathcal{D}'_{s_1}$ à $\mathcal{E}'_x$ dans $\mathcal{E}_x$

$$\mathcal{E}_x = \mathcal{E}'_x \overset{\perp}{\oplus} \mathcal{D}'_{s_1}. \tag{6.10}$$

Remarquons que la plus petite valeur propre κ_{N_s} peut être interprétée comme la puissance moyenne de la première source $s_1(t)$ pendant $\mathcal{T}_1$ et devra donc être proche de zéro. Finalement, la première ligne $G_{1,:}$ de G, permettant d'extraire $s_1(t)$ des observations $\mathbf{x}(t)$, est définie par

$$G_{1,:} = \mathbf{g}^+_{N_s}. \tag{6.11}$$

Ainsi, pour l'ensemble $\mathcal{T}$ de tous les indices temporels (incluant maintenant les indices où $s_1(t)$ est active), nous pouvons extraire la première source $s_1(t)$ grâce à :

$$y_1(t) = G_{1,:}\,\mathbf{x}(t) = C_{1,1}s_1(t) \tag{6.12}$$

où $C_{1,1} = G_{1,:}\,H_{:,1}$. Notons que ce gain $C_{1,1}$ peut s'interpréter comme une distorsion non contrôlée puisque $\mathcal{D}_1$ est, *a priori*, un supplémentaire de $\mathcal{E}'_x$ mais pas nécessairement le supplémentaire orthogonal de $\mathcal{E}'_x$ dans $\mathcal{E}_x$. Comme expliqué ci-après (paragraphe 6.2), dans le cas convolutif, cette distorsion peut altérer de façon considérable l'estimation de la source. C'est pourquoi nous allons montrer maintenant que l'on peut fixer le gain à $H_{1,1}$ au lieu de $C_{1,1}$, *i.e.* nous pouvons trouver un scalaire complexe λ tel que

$$\hat{s}_1(t) = \lambda\, y_1(t) = H_{1,1}\, s_1(t). \tag{6.13}$$

Ainsi λ est donné grâce à (6.12) par

$$\lambda = \frac{H_{1,1}}{\sum_{i=1}^{N_s} G_{1,i}\, H_{i,1}} = \frac{1}{G_{1,1} + \sum_{i>1} G_{1,i}\, H_{i,1}/H_{1,1}} \tag{6.14}$$

où $\forall\, i$, les coefficients $G_{1,i}$ sont connus et l'ensemble $\{H_{i,1}/H_{1,1}\}_i$ est à estimer. Pour estimer ces coefficients, nous proposons une procédure fondée sur l'annulation de la contribution de la source estimée $\hat{s}_1(t)$ dans les différents mélanges $x_i(t)$. Posons

$$\epsilon_i(\beta_i) = \mathrm{E}\left[|x_i(t) - \beta_i\, y_1(t)|^2\right]. \tag{6.15}$$

Puisque les sources sont indépendantes, à partir de (6.1) et (6.12) nous obtenons

$$\epsilon_i(\beta_i) = \mathrm{E}\left[|(H_{i,1} - \beta_i\, C_{1,1}) s_1(t)|^2\right] + \sum_{j>1} \mathrm{E}\left[|H_{i,j}\, s_j(t)|^2\right]. \tag{6.16}$$

Ainsi, $\forall\, \beta_i$, $\epsilon_i(\beta_i)$ a pour borne inférieure $\sum_{j>1} \mathrm{E}\left[|H_{i,j}\, s_j(t)|^2\right]$ et cette borne est atteinte lorsque $\beta_i = H_{i,1}/C_{1,1}$. Soit $\hat{\beta}_i$ l'estimation optimale de β_i au sens du minimum de l'erreur quadratique moyenne (*i.e.* minimisation de $\epsilon_i(\beta_i)$). $\hat{\beta}_i$ est obtenue de façon classique par

$$\hat{\beta}_i = \arg\min_{\beta_i} \epsilon_i(\beta_i) = \frac{\mathrm{E}[x_i^*(t)\, y_1(t)]}{\mathrm{E}[|y_1(t)|^2]}. \tag{6.17}$$

En pratique l'espérance est remplacée par la moyenne temporelle et $\hat{\beta}_i$ est donné par

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T} x_i^*(t) y_1(t)}{\sum_{t=1}^{T} |y_1(t)|^2}. \tag{6.18}$$

Ainsi, λ est donné par (6.14) où $H_{i,1}/H_{1,1}$ est remplacé par $\hat{\beta}_i/\hat{\beta}_1$:

$$\hat{\lambda} = \frac{1}{G_{1,1} + \sum_{i>1} G_{1,i}\, \hat{\beta}_i/\hat{\beta}_1}. \tag{6.19}$$

Notons que l'on a utilisé le rapport $H_{i,1}/H_{1,1}$ plutôt que $H_{i,1}$ directement puisque β_i est égal à $H_{i,1}$ au coefficient inconnu $C_{1,1}$ près. Finalement, la source $s_1(t)$ est estimée par

$$\hat{s}_1(t) = \hat{\lambda}\, G_{1,:}\, \mathbf{x}(t) = \hat{H}_{1,1}\, s_1(t). \tag{6.20}$$

Ce principe est résumé dans l'algorithme 4.

6.2 Cas des mélanges convolutifs complexes

Nous allons maintenant étendre le principe que nous venons de présenter au cas des mélanges convolutifs.

Soient maintenant N_s sources centrées $\mathbf{s}(t) = [s_1(t), \cdots, s_{N_s}(t)]^T$ et autant d'observations $\mathbf{x}(t) = [x_1(t), \cdots, x_{N_s}(t)]^T$ ($N_o = N_s$) obtenues par un processus de mélanges convolutif. Nous rappelons que la séparation de sources de mélanges convolutifs est généralement considérée dans le domaine fréquentiel où l'unique problème

Algorithme 4 Extraction directe par la parcimonie pour les mélanges instantanés.

/Estimation des moments de silence de $s_1(t)$/
Estimer $\mathcal{T}_1$ grâce au détecteur d'activité vocale visuel (chapitre 4)

/Estimation de $G_{1,:}$/
Calculer $C_{\mathbf{xx}}(\mathcal{T}_1) = \mathrm{E}_{t\in\mathcal{T}_1}\left[\mathbf{x}(t)\mathbf{x}^+(t)\right]$
Faire la décomposition en valeurs propres de $C_{\mathbf{xx}}(\mathcal{T}_1)$
Selectionner $\mathbf{g}_{N_s}$ le vecteur propre associé avec la plus petite valeur propre κ_{N_s}
$G_{1,:} \Leftarrow \mathbf{g_{N_s}}^+$

/Estimation de λ pour fixer le gain/
Estimer β_i avec (6.18)
λ est donné par (6.19)

Estimer la source $s_1(t)$ grâce à (6.20)

convolutif devient N_f (le nombre de fréquences de calcul) problèmes complexes instantanés. A chaque fréquence f, les équations de mélange et séparation deviennent donc

$$X_m(t,f) = \sum_{n=1}^{N_s} H_{m,n}(f)S_n(t,f) \tag{6.21}$$

$$Y_n(t,f) = \sum_{m=1}^{N_s} G_{n,m}(f)X_m(t,f) \tag{6.22}$$

où $S_n(t,f)$, $X_m(t,f)$ et $Y_n(t,f)$ sont les transformées de Fourier à court terme (TFCT) de respectivement $s_n(t)$, $x_m(t)$ et $y_n(t)$. $H_{m,n}(f)$ et $G_{n,m}(f)$ sont les réponses fréquentielles des filtres de mélanges $H(f)$ et de séparation $G(f)$, respectivement. Puisque la fonction de mélange est supposée stationnaire, $H(f)$ et $G(f)$ ne dépendent pas du temps, tandis que les signaux (*i.e.* sources, observations) peuvent être non stationnaires. Dans le domaine fréquentiel, le but de la séparation de sources est donc d'estimer, à chaque fréquence f, la matrice de filtres de séparation $G(f)$. Ceci peut être fait grâce à la méthode géométrique proposée au paragraphe 6.1. En effet, à chaque fréquence f, les équations (6.21) et (6.22) peuvent être interprétées comme des mélanges instantanés de sources complexes. Ainsi, $G_{1,:}(f)$ est le transconjugué du vecteur propre associé à la plus petite valeur propre de la matrice de covariance $C_{\mathbf{xx}}(\mathcal{T}_1, f) = \mathrm{E}_{t\in\mathcal{T}_1}\left[\mathbf{X}(t,f)\mathbf{X}^+(t,f)\right]$. De plus, $\beta_i(f)$ dépend de la fréquence et peut être estimé par

$$\hat{\beta}_i(f) = \frac{\sum_{t=1}^{T} X_i^*(t,f)Y_1(t,f)}{\sum_{t=1}^{T} |Y_1(t,f)|^2}. \tag{6.23}$$

donc $\lambda(f)$ est maintenant estimé par

$$\hat{\lambda}(f) = \frac{1}{G_{1,1}(f) + \sum_{i>1} G_{1,i}(f)\,\hat{\beta}_i(f)/\hat{\beta}_1(f)}. \tag{6.24}$$

Finalement, pour estimer les sources, nous calculons la réponse impulsionelle (RI) des filtres de séparation par transformée de Fourier inverse de $\hat{\lambda}(f)\,G_{1,:}(f)$. La source

$S_1(t,f)$ est alors estimée par

$$\hat{s}_1(t) = \mathrm{TF}^{-1}\big[\hat{\lambda}(f)\, G_{1,:}(f)\big] * \mathbf{x}(t), \tag{6.25}$$

où $\mathrm{TF}[\cdot]^{-1}$ est l'opérateur transformée de Fourier inverse.

Notez que dans le cas convolutif, si la régularisation du gain $\lambda(f)$ n'est pas assurée, la source $s_1(t)$ est estimée à un filtre inconnu qui peut altérer perceptuellement l'estimation de la source de façon très importante . Au contraire, effectuer cette régularisation du gain assure que la source $s_1(t)$ est estimée au filtre $H_{1,1}(t)$ près, filtre qui correspond au filtre "canal+capteur" de la première observation. On a donc une situation équivalente à celle où l'estimation de la source $s_1(t)$ correspond à ce qu'aurait enregistré le premier capteur en admettant qu'aucune des autres sources ne soient présentes. La méthode complète est résumée dans l'algorithme 5.

Algorithme 5 Séparation géométrique de mélanges convolutifs

/Estimation des moments de silence de $s_1(t)$/
Estimer $\mathcal{T}_1$ grâce au détecteur d'activité vocale visuel (chapitre 4)

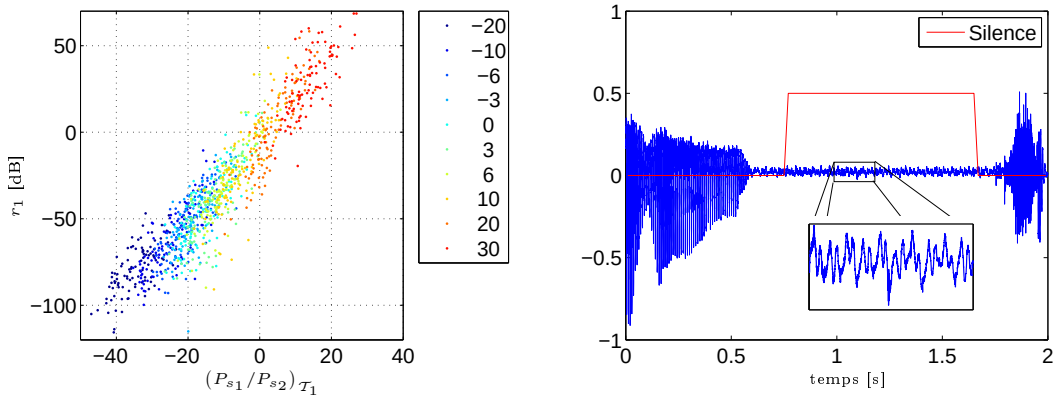
Calculer les TFCT des observations $x_m(t)$ pour obtenir $X_m(t,f)$

Pour toutes les fréquences f **faire**
/Estimation de $G_{1,:}(f)$/
Calculer $C_{\mathbf{xx}}(\mathcal{T}_1, f) = \mathrm{E}_{t\in\mathcal{T}_1}\big[\mathbf{X}(t,f)\mathbf{X}^{+}(t,f)\big]$ avec $t \in \mathcal{T}_1$
Effectuer la décomposition en valeur propres de $C_{\mathbf{xx}}(\mathcal{T}_1, f)$
Sélectionner $\mathbf{g}_{N_s}(f)$ le vecteur propre associé à la plus petite valeur propre $\kappa_{N_s}(f)$
$G_{1,:}(f) \Leftarrow \mathbf{g}^{+}_{N_s}(f)$

/Estimation de $\lambda(f)$ pour régulariser le facteur d'échelle/
Estimer $\beta_i(f)$ avec (6.23)
$\hat{\lambda}(f)$ est donné par (6.24)
Fin boucle
/Estimation des réponses impulsionnelles des filtres de séparation/
Calculer la TF inverse de $\Lambda(f)\, G_{1,:}(f)$ pour obtenir la RI du filtre de séparation

/Estimation des sources/
Estimer la source $s_1(t)$ grâce à (6.25)

Remarquons que la décomposition en valeurs propres de la matrice de covariance $C_{\mathbf{xx}}(\mathcal{T}_1, f)$ des observations estimée pendant les silences de la source $s_1(t)$ fournit N_s valeurs propres correspondant à la répartition des puissances moyennes pendant $\mathcal{T}_1$ dans l'espace des mélanges $\mathcal{E}_x$. Le vecteur propre $\mathbf{g}_{N_s}(f)$ associé à la plus petite valeur propre $\kappa_{N_s}(f)$ correspond à la direction représentant le moins de puissance moyenne dans les mélanges durant les périodes de silence de $s_1(t)$. Projeter les observations sur la direction du vecteur propre associé à la plus petite valeur propre pour extraire la source $s_1(t)$ revient donc à projeter les observations sur la direction représentant le moins de puissance. Ceci est naturel puisque pendant $\mathcal{T}_1$ la source $s_1(t)$ s'annule.



(a) Indice de performance (b) Source $s_1(t)$

FIG. 6.3 – Performances dans le cas instantané. Figure 6.3(a) : indice de performance (5.32) en fonction du rapport des puissances des sources pendant les moments de silence. Ce rapport est exprimé en dB. La légende indique le rapport signal sur bruit en entrée en dB. Figure 6.3(b) : source $s_1(t)$ avec l'indexation manuelle *silence* en rouge.

6.3 Résultats expérimentaux

Dans ce paragraphe, nous présentons les résultats de l'extraction d'une source de parole par la méthode que nous venons de présenter tout d'abord dans le cas de mélanges instantanés, puis dans le cas réaliste de mélanges convolutifs.

Dans ces expériences, la source d'intérêt est issue du corpus de "Grenoble", nous rappelons que ce corpus a été enregistré dans des conditions de dialogue spontané (*cf.* chapitre 4). L'oracle, qui nous fournit l'ensemble des indices temporels $\mathcal{T}_1$ où la source d'intérêt $s_1(t)$ est absente, est obtenu soit par le détecteur de silence visuel du chapitre 4, soit par l'indexation manuelle suivant le type d'expérience.

6.3.1 Cas des mélanges instantanés

Dans cette série d'expériences, nous présentons les performances obtenues par le principe de séparation que nous avons introduit dans le cas de mélanges instantanés. Pour cela, à chaque expérience, la matrice de mélange est choisie de la forme

$$H = \begin{pmatrix} \cos\theta_1 & \cos\theta_2 \\ \sin\theta_1 & \sin\theta_2 \end{pmatrix}$$

où θ_1 et θ_2 sont déterminés aléatoirement. La source $s_1(t)$ (resp. $s_2(t)$) correspond à 10 secondes de signal choisies aléatoirement parmi l'ensemble du corpus "Grenoble" (resp. de phrases). Le rapport signal sur bruit en entrée, défini comme le rapport entre les puissances moyennes des deux sources pendant les moments indexés manuellement *non silence*, varie de -20dB à 30dB. Pour chacun de ces rapports signal sur bruit, 100 configurations de mélange ont été réalisées. De façon à estimer la première ligne $G_{1,:}$, de la matrice de séparation G, nous appliquons le principe donné

dans l'algorithme 4 où la détermination de l'ensemble des moments de silence $\mathcal{T}_1$ est ici remplacée par l'indexation manuelle des silences, que nous avons fait au préalable. Les résultats, en terme d'indice de performance, sont présentés à la figure 6.3. L'indice de performance de chacune des configurations est donné en fonction du rapport des puissances moyennes des deux sources, P_{s_1} et P_{s_2} calculées pendant les instants indexés *silence* $\mathcal{T}_1$, rapport que l'on note $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$. Comme on peut le voir à la figure 6.3(a), plus ce rapport est petit, meilleure est l'estimation de la première ligne de la matrice de séparation : *i.e.* plus l'indice de performance (5.32) est petit. Il est intéressant de noter à ce sujet que l'on obtient une forte corrélation, 93%, entre l'indice de performance et le rapport des puissances moyennes pendant les silences. Ceci s'explique par le fait que notre principe revient à projeter les mélanges dans la direction représentant le moins de puissance pendant les moments où la source $s_1(t)$ est absente (*i.e.* $s_1(t) = 0$). Ainsi, si $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ est petit alors cette direction correspond effectivement à l'hyperplan $\mathcal{D}'_{s_1}$ orthogonal à $\mathcal{E}'_x$. Donc l'indice de performance est petit, traduisant le fait que la ligne $G_{1,:}$ est orthogonale à toutes les colonnes $H_{:,i}$, pour $i \neq 1$. De plus on constate que plus le rapport signal sur bruit en entrée est important, moins bonne est l'estimation de la ligne $G_{1,:}$. Ceci peut s'expliquer par le fait que pendant les instants indexés *silence*, le signal $s_1(t)$ n'est pas rigoureusement nul comme le montre la figure 6.3(b) où est tracé une portion du signal $s_1(t)$ ainsi qu'un zoom de celui-ci pendant du silence. Ainsi, si le rapport signal sur bruit en entrée augmente, le rapport $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ augmente aussi ce qui diminue les performances de l'extraction.

Remarquons de plus que si le rapport $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ devient supérieur à 1, ou positif si celui-ci est exprimé en dB, alors l'extraction entraîne un indice de performance supérieur à 1 et donc un gain, défini par (5.35), inférieur à 1 correspondant à une dégradation du rapport signal sur interférence (5.34) pour la source à extraire. Ceci amène deux conclusions sur la robustesse de notre principe vis-à-vis des erreurs d'indexation des moments de silence d'une part et vis-à-vis d'un bruit additif d'autre part. En effet, tant que le détecteur de silence que nous utilisons pour la détermination de $\mathcal{T}_1$ ne fait pas trop d'erreur (*i.e.* assure que le rapport $\left(P_{s_1}/P_{s_2}\right)_{\mathcal{T}_1}$ est très inférieur à 1) alors l'estimation de la ligne $G_{1,:}$, permettant l'extraction de la source d'intérêt, est correcte (*i.e.* est telle que l'indice de performance est très inférieur à 1). D'autre part, en présence d'un bruit additif[1] $\mathbf{b}(t)$ sur les mélanges $\mathbf{x}(t) = H\mathbf{s}(t) + \mathbf{b}(t)$, la répartition de la puissance dans l'espace $\mathcal{E}_x$, pendant les instants de silence $\mathcal{T}_1$, est modifiée de la même façon que la nouvelle matrice de covariance des observations $C_{\mathbf{x},\mathbf{x}}(\mathcal{T}_1) = HC_{\mathbf{s},\mathbf{s}}(\mathcal{T}_1)H^+ + C_{\mathbf{b},\mathbf{b}}(\mathcal{T}_1)$, en supposant que le bruit est indépendant des sources. La décomposition en valeurs propres de cette matrice de covariance calculée pendant les silences de la source $s_1(t)$ fournit N_s valeurs propres et de façon classique on attribue les plus grandes au sous-espace signal tandis que les plus petites correspondent au sous-espace bruit. Ainsi, de façon qualitative, tant que le vecteur propre associé à la plus petite valeur propre de $C_{\mathbf{x},\mathbf{x}}(\mathcal{T}_1)$ est orthogonal au sous-espace signal (*i.e.* est orthogonal à toutes les colonnes $H_{:,i}$, pour $i \neq 1$) alors l'extraction sera correcte. En revanche, si le vecteur propre associé à la plus petite valeur propre de $C_{\mathbf{x},\mathbf{x}}(\mathcal{T}_1)$ est orthogonal au sous-espace bruit alors l'extraction ne pourra donner de

[1] Ce bruit additif peut par exemple modéliser le fait que le mélange est en réalité sous-déterminé.

bonnes performances. Cette dernière situation peut se produire quand la puissance du bruit devient grande devant la puissance des autres sources pendant les instants de silence de la source $s_1(t)$: l'hypothèse d'attribuer les plus grandes valeurs propres au sous-espace signal n'est alors plus vérifiée.

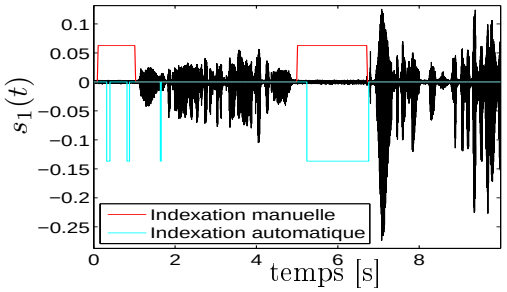
Nous présentons maintenant à la figure 6.4 un exemple d'extraction d'une source de parole à partir d'un mélange instantané où la matrice de mélange H est de la forme

$$H = \begin{pmatrix} \cos\theta_1 & \cos\theta_2 \\ \sin\theta_1 & \sin\theta_2 \end{pmatrix}$$

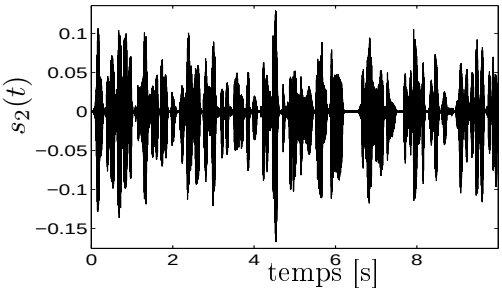
où $\theta_1 = 45°$ et $\theta_2 = 30°$. Dans cet exemple, dix secondes de signal sont utilisés pour estimer la première ligne de la matrice de séparation. Les figures 6.4(a) et 6.4(b) montrent les deux sources qui ont la même puissance moyenne. Les figures 6.4(d) et 6.4(f) montrent les deux mélanges. L'évolution des paramètres labiaux (largeur et hauteur internes) est représentée à la figure 6.4(c) où l'on a aussi tracé le paramètre vidéo intégré $\Pi(t)$ utilisé pour la détection visuelle des silences (le trait horizontal rouge correspond au seuil choisi). L'estimation de la première ligne de la matrice de séparation est correcte puisque le gain, défini par (5.35), est égal à 45 dB. Ces bonnes performances sont confirmées par le tracé de l'estimation de la première source $\hat{s}_1(t)$ à la figure 6.4(e). Sur la figure 6.4(g), nous avons représenté la distribution conjointe des deux sources $\mathcal{E}_s$ ainsi que les directions des espaces (orthogonaux) qu'elles engendrent $\mathcal{E}_{s_1}$ et $\mathcal{E}_{s_2}$ respectivement. La figure 6.4(h) montre la distribution conjointe des deux mélanges avec les directions des espaces engendrés par chacune des sources $\mathcal{D}_{s_1}$ et $\mathcal{D}_{s_2}$. Les points verts correspondent à l'espace $\hat{\mathcal{E}}'_x$, c'est-à-dire aux couples $\big(x_1(t), x_2(t)\big)$ pendant les instants détectés comme *silence* par le détecteur de silence visuel (*i.e.* $t \in \mathcal{T}_1$). La droite bleue correspond à l'estimation du supplémentaire orthogonal à $\hat{\mathcal{E}}'_x$, c'est donc la direction de $\mathcal{D}'_{s_1}$ qui sert à l'extraction de $s_1(t)$. Comme on peut le voir sur cette figure, $\mathcal{D}'_{s_1}$ est bien orthogonal à $\mathcal{D}_{s_2}$, ce qui est confirmé par l'indice de performance, défini par (5.32), égal à -47dB.

6.3.2 Cas des mélanges convolutifs

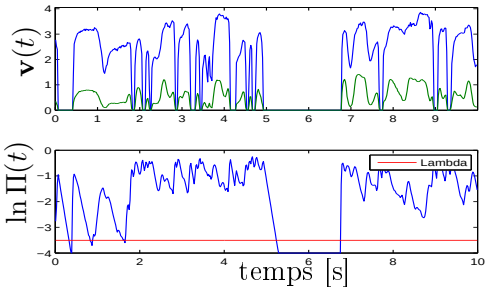
Nous présentons à la figure 6.5 les performances obtenues par le principe d'extraction d'une source parole par la parcimonie dans le cas des mélanges convolutifs. Pour chaque configuration de la matrice de filtre ($N_s \times N_s$) et du rapport signal sur interférence en entrée (défini comme la moyenne des rapports signal sur interférence pour chaque mélange), les simulations ont été répétées 50 fois. La source d'intérêt $s_1(t)$ est une section de 20 secondes choisies aléatoirement parmi la totalité du corpus de "Grenoble", les autres sources sont de la parole issue du corpus des phrases. L'indexation des trames de silence $\mathcal{T}_1$ est donnée par le détecteur de silence visuel du chapitre 4. On peut voir que, dans les deux cas, l'allure des courbes est la même : d'un RSI faible (-20dB) à un RSI moyen (0dB), les gains sont relativement constant à une valeur élevée démontrant ainsi l'efficacité de la méthode : nous obtenons des gains d'environ 18–19dB dans le cas (2×2) et d'environ 17–18dB dans le cas (3×3). Ensuite, les gains diminuent jusqu'à 11dB pour le cas (2×2) et 8dB dans le cas (3×3) pour de fort RSI. Il est intéressant de remarquer que, comme pour le cas instantané, le gain peut être négatif pour de fort RSI (par exemple dans le cas (2×2)



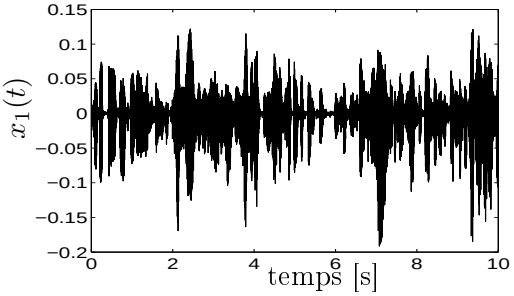
(a) Source 1 : $s_1(t)$



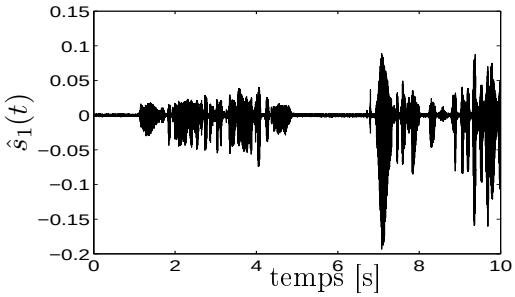
(b) Source 2 : $s_2(t)$



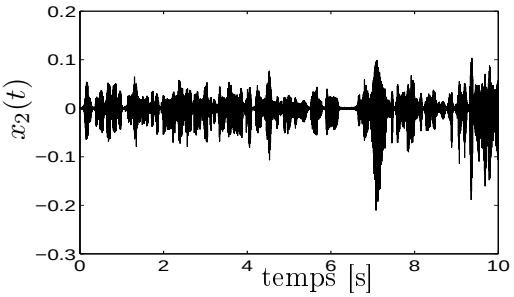
(c) Paramètres vidéo



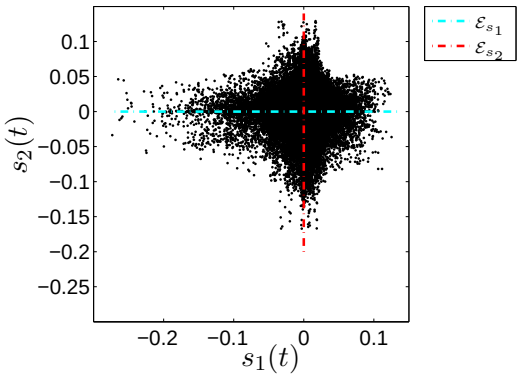
(d) Mélange 1 : $x_1(t)$



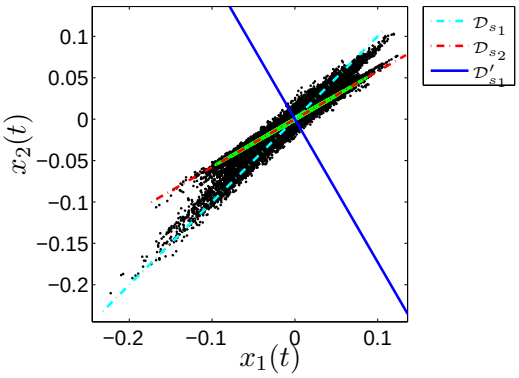
(e) Source estimée 1 : $\hat{s}_1(t)$



(f) Mélange 2 : $x_2(t)$

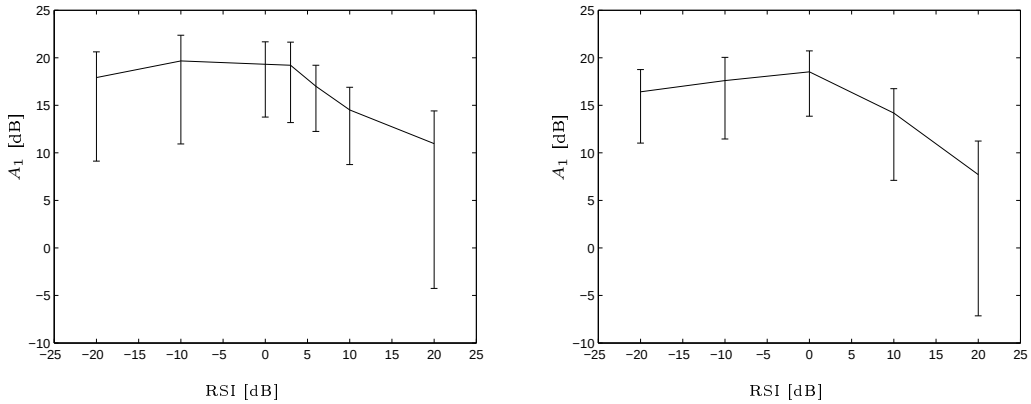


(g) Espace des sources $\mathcal{E}_s$



(h) Espace des mélanges $\mathcal{E}_x$

FIG. 6.4 – Exemple d'extraction directe par la parcimonie pour un mélange instantané.

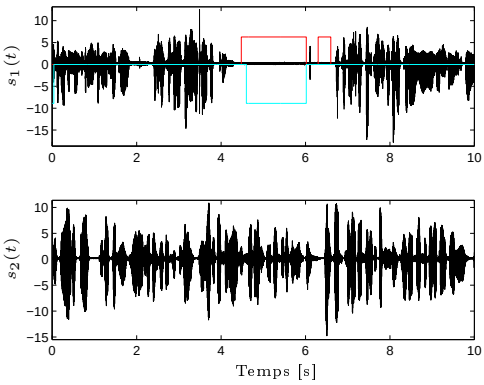


(a) Cas (2×2) (b) Cas (3×3)

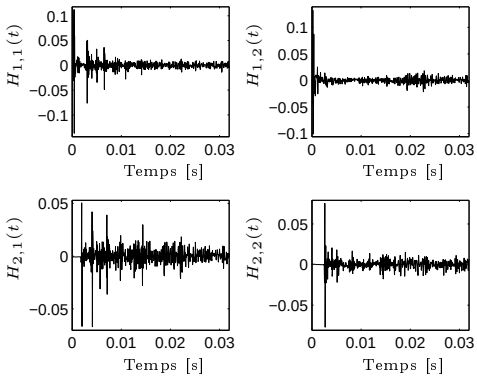
FIG. 6.5 – Performances dans le cas convolutif. Gain A_1 (5.35) en fonction du RSI d'entrée moyen dans le cas (2×2) (figure 6.5(a)) et (3×3) (figure 6.5(b)). Sont tracés les valeurs moyennes et les écarts types en dB.

pour un RSI de 20dB). Ce cas se produit si le taux entre les fausses détections de silence et les bonnes détection de silence est important et que le RSI en entrée est grand : le vecteur propre associée à la plus petite valeur propre n'est alors pas nécessairement orthogonal à toutes les colonnes $H_{:,i}(f)$, pour $i \neq 1$. En revanche, pour de faibles RSI en entrée, même si ce taux est important, son influence est moindre puisque même en cas de fausse détection de silence, les valeurs prises par $S_1(t, f)$ sont petites devant celles prises par les autres sources.

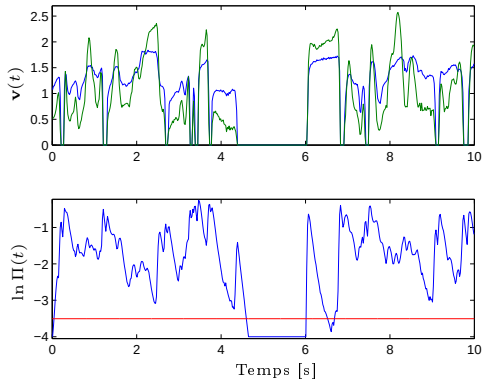
La figure 6.6 présente un résultat typique dans le cas (2×2) de deux sources et deux capteurs avec approximativement un rapport signal sur interférence (RSI défini par (5.34)) pour la source $s_1(t)$ égal à 0 pour les deux capteurs. Les deux sources de paroles sont tracées à la figure 6.6(a). Nous avons représenté à la figure 6.6(c) l'évolution des paramètres labiaux (largeur et hauteur internes des lèvres) ainsi que le logarithme du paramètre vidéo intégré $\Pi(t)$ (le trait rouge correspond au seuil utilisé pour le détecteur de silence visuel). Dans cet exemple, le coefficient d'intégration α est égal à $0.82 = \exp(-1/\tau)$ avec $\tau = 5$ et le nombre minimal de trames de silence que l'on s'autorise à détecter est de 20. Les deux observations $x_1(t)$ et $x_2(t)$ obtenues à partir des sources et des filtres de mélanges, dont les réponses impulsionnelles sont tracées à la figure 6.6(b), sont reportées sur la figure 6.6(d). Le résultat de l'extraction de la source $s_1(t)$ par la méthode proposée dans ce chapitre est donnée à la figure 6.6(e) où l'on a aussi reporté la meilleure estimation possible de la source d'intérêt donnée par $H_{1,1}(t) * s_1(t)$. Dans cet exemple, le gain est de 18dB tandis que le rapport signal sur interférence en sortie est égal à 19dB. On peut voir que l'extraction de la première source est relativement bien effectuée. Ceci est confirmé par l'indice de performance $r_1(f)$ proche de 0 pour la grande majorité des fréquences (*cf.* figure 6.6(f)). De plus les valeurs prises par la réponse impulsionnelle $C_{1,2}(t)$ sont très inférieures à celles prises par $C_{1,1}(t)$ ce qui confirme que les composantes fréquentielles, pour lesquelles l'indice de performance $r_1(f) > 1$, n'ont que très peu d'influence sur l'estimation des filtres de séparation.



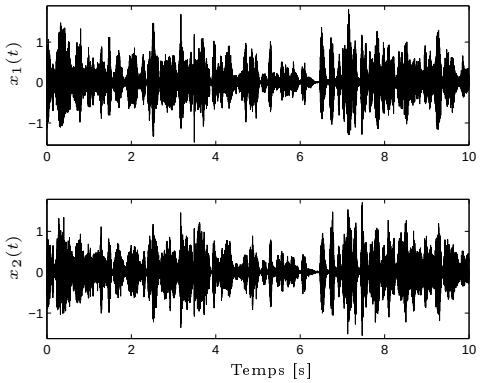
(a) Sources



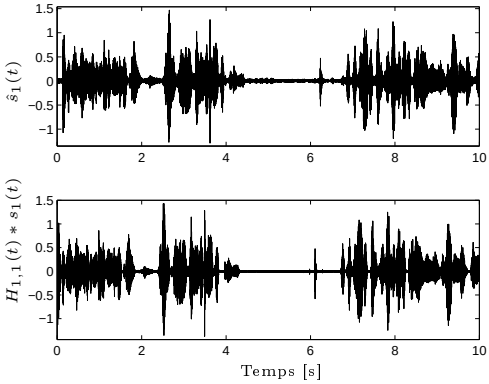
(b) Filtres de mélange



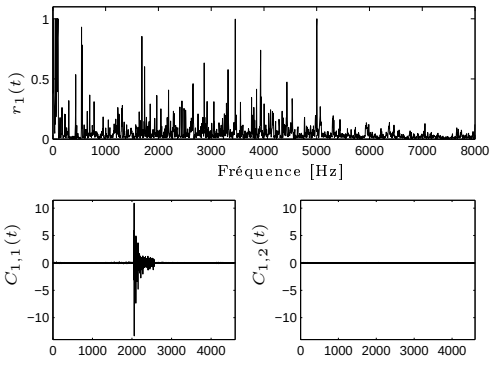
(c) Paramètres vidéo



(d) Mélanges



(e) Estimation



(f) Performances

FIG. 6.6 – Exemple d'extraction directe par la parcimonie dans le cas convolutif de deux sources et deux capteurs.

6.4 Conclusion

Dans ce chapitre, nous avons proposé une méthode efficace qui exploite la parcimonie de la parole. La modalité visuelle est utilisée pour une détection d'activité vocale, tandis que la modalité acoustique est utilisée pour l'estimation des matrices de séparation en exploitant les moments de silence du locuteur à extraire. La méthode géométrique que nous proposons est fondée sur la parcimonie du signal de parole spontanée : quand la source à extraire s'annule, la puissance de la source estimée correspondante est minimisée grâce aux filtres de séparation. Les expériences dans le cas de deux sources et de deux capteurs ou trois sources et trois capteurs ont montré que la méthode est performante. De plus, notons le faible coût calculatoire de cette méthode puisque, comparée aux méthodes présentées au chapitre précédent fondées sur une diagonalisation conjointe de plusieurs matrices, elle ne nécessite que la diagonalisation d'une seule matrice de covariance. Enfin, cette méthode est peu dépendante du nombre de sources, tant que la matrice de mélange est carrée, et ne fait pas d'hypothèse sur la nature des sources concurrentes de la source à extraire.

Conclusion générale et perspectives

Dans ce travail de thèse, nous avons présenté une étude pour l'extraction d'un locuteur à partir de mélanges de type convolutifs en exploitant la bimodalité audio/vidéo de la parole. Pour cela, nous sommes partis de la modélisation de la bimodalité de la parole pour l'inclure dans des systèmes de séparation aveugle de sources avant de proposer une nouvelle méthode spécifiquement adaptée à l'extraction d'un locuteur. Plutôt que de proposer une conclusion générale sur l'ensemble de notre travail, nous préférons reprendre les deux grands aspects abordés en précisant à chaque fois les perspectives qui y sont liées.

Modélisation de la bi-modalité de la parole

Tout d'abord nous avons introduit un nouveau modèle statistique multi-noyaux spécifiquement adapté au logarithme des coefficients de la TFCT de la parole continue. Chaque noyau LogRayleigh est construit de telle sorte qu'il modélise au mieux un unique son de parole avec une seule matrice de localisation. Nous avons ensuite étendu ce modèle audio de façon à modéliser la parole audiovisuelle : chaque noyau permet alors d'associer à une enveloppe spectrale d'un son la forme des lèvres correspondante.

Notre seconde contribution à la modélisation de la bimodalité de la parole s'est portée sur la détection d'activité vocale. Jusque-là abordée de façon purement acoustique, nous l'avons étendue au cas audiovisuel en exploitant le modèle audiovisuel précédent. Ensuite, nous avons proposé un nouveau détecteur de silence purement visuel qui exploite l'absence de mouvement des lèvres pendant le silence. Un tel détecteur de silence présente l'avantage d'être robuste à tout type de bruit acoustique, notamment si celui-ci est fortement non stationnaire. Pour tester ses performances, nous avons conçu le protocole et enregistré une nouvelle base de données en plusieurs langues permettant d'associer des signaux audio d'un locuteur avec les vues des lèvres, soit de face soit de profil.

Une première perspective à ce travail serait d'exploiter la totalité de la base de données que nous avons enregistrée de façon à tester notre détecteur de silence visuel sur un plus grand nombre de sujets pour étudier sa dépendance vis-à-vis du locuteur. Il serait également intéressant d'étudier la robustesse du détecteur d'activité vocale visuel vis-à-vis du bruit vidéo (conditions d'éclairage, ombres, *etc.*) Il serait également envisageable d'utiliser le détecteur d'activité vocale visuel comme un *a priori* pour un détecteur d'activité vocale acoustique. Cela permettrait d'allier

la simplicité de la détection d'activité visuelle à celle de la détection d'activité vocale acoustique qui exploite un modèle global à long terme.

Extraction d'une source de parole audiovisuelle

Nous avons ensuite exploité nos modélisations de la bimodalité de la parole pour extraire le signal d'un locuteur particulier. Dans un premier temps, ces modélisations nous ont permis de résoudre le problème des indéterminations liées aux méthodes aveugles de séparation de sources fondées sur l'indépendance. Pour cela, nous avons utilisé tout d'abord le modèle audiovisuel de façon à reconstruire la source la plus cohérente possible avec le signal vidéo du locuteur. Bien qu'efficace, cette méthode présente l'inconvénient d'être fortement dépendante du locuteur de part la modélisation audiovisuelle liant forme des lèvres aux paramètres spectraux du son prononcé. Ensuite, pour obtenir un algorithme moins coûteux en temps de calcul, nous avons utilisé le détecteur de silence purement visuel, dont la dépendance vis-à-vis du locuteur ne s'exprime plus que dans un simple seuil. Comme l'on montré nos expériences, cette méthode est relativement robuste aux erreurs de détection des silences.

Nous avons finalement présenté une méthode fondée sur l'emploi direct de la bimodalité de la parole. Celle-ci exploite les moments de silence du locuteur d'intérêt de façon à pouvoir l'extraire quand il parle. Cette dernière méthode présente en outre l'avantage de résoudre intrinsèquement le problème des permutations. Elle est de plus peu coûteuse en temps de calcul comparée aux méthodes précédentes car elle ne nécessite que la diagonalisation d'une matrice et non pas une diagonalisation conjointe d'un ensemble de matrices.

Comme perspectives à ces travaux, l'emploi du modèle audiovisuel pour résoudre les permutations s'est montré performant mais coûteux en temps de calcul. Ainsi, il serait intéressant de chercher d'autres algorithmes comme par exemple les algorithmes génétiques. Proposer un modèle exploitant le lien temporel du signal audiovisuel pourrait peut-être améliorer encore ses performances. Dans notre travail, l'estimation des filtres de séparation a été faite hors-ligne. Une piste serait de proposer des algorithmes en ligne de façon à pouvoir s'adapter à des changements des conditions de mélanges. Il serait également intéressant de faire une analyse plus fine de la robustesse de notre méthode à un bruit additif et de la complexité de nos algorithmes. Finalement, il est intéressant de constater que les méthodes que nous avons proposées exploitant le détecteur de silence visuel peuvent être utilisées pour d'autres applications pour peu que l'on puisse construire un oracle indiquant l'absence de la source à extraire.

La séparation de sources est un domaine de recherche attractif dont les approches actuelles favorisent l'exploitation d'informations *a priori* très variées. En particulier, la bimodalité audio/vidéo de la parole conduit à des méthodes performantes. Nous pensons que la séparation de sources audiovisuelle mériterait d'être beaucoup plus largement explorée.

Annexes

Annexe A

Distribution de LogRayleigh

Dans cette annexe, nous étudions la distribution suivie par le logarithme du module d'une variable aléatoire complexe gaussienne [89, 103, 114] dans le cas circulaire puis dans le cas non circulaire.

A.1 Distribution de LogRayleigh circulaire

Soit X une variable aléatoire centrée complexe gaussienne circulaire de variance σ^2 : $X \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. La densité de probabilité de X est alors donnée par [89, 103]

$$p_X(x) = \left[\pi\sigma^2\right]^{-1} \exp\left[-\frac{|x|^2}{\sigma^2}\right].$$

Cette équation suppose que les parties réelles et imaginaires de X sont décorrélées et de même variance égale à $\sigma^2/2$:

$$\begin{cases} \Re\{X\} & \sim & \mathcal{N}_{\mathbb{R}}\left(0, \frac{\sigma^2}{2}\right) \\ \Im\{X\} & \sim & \mathcal{N}_{\mathbb{R}}\left(0, \frac{\sigma^2}{2}\right) \end{cases}$$

où $\Re\{\cdot\}$ et $\Im\{\cdot\}$ sont respectivement les opérateurs parties réelle et imaginaire. Il est bien connu [92] que le module $Y = |X| = \sqrt{\Re\{X\}^2 + \Im\{X\}^2}$ de X suivant une loi de Rayleigh de paramètre $\sigma^2/2$: $Y \sim \text{Ray}(\sigma^2/2)$. La densité de probabilité d'une loi de Rayleigh de paramètre β^2 est donnée par

$$p_Y(y) = \begin{cases} \frac{y}{\beta^2} \exp\left(-\frac{y^2}{2\beta^2}\right) & \text{pour } y \geq 0, \\ 0 & \text{pour } y < 0. \end{cases}$$

Ainsi, soit $Z = \ln Y$ le logarithme du module de X. La distribution de Z peut être obtenue en utilisant la propriété A.1.

Propriété A.1 (Changement de variable)
Soient U et V deux variables aléatoires telles que $V = h(U)$ où $h(\cdot)$ est une fonction inversible. Soient $p_U(\cdot)$ et $p_V(\cdot)$ les densités de probabilité de U et V respectivement. On a alors

$$p_V(v) = \left(\frac{1}{\left|\frac{\partial h}{\partial u}\right|} p_U(u)\right)_{v=h(u)}. \tag{A.1}$$

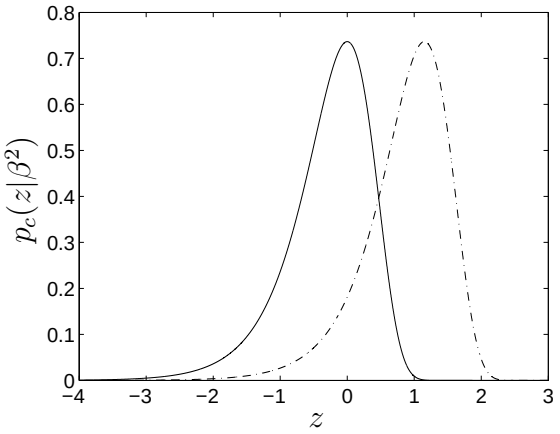


FIG. A.1 – Densité de probabilité d'une loi LogRayleigh circulaire de paramètre de localisation $\beta^2 = 1$ (ligne continue) et $\beta^2 = 10$ (ligne tiretée).

Ainsi, si $h(\cdot) = \ln(\cdot)$ et $p_U(\cdot)$ est la densité de probabilité Rayleigh de paramètre β^2 alors la distribution de $V = \ln U$ vérifie

$$\forall v \in \mathbb{R}, \quad p_V(v) = p_c(v|\beta^2) = \frac{(e^v)^2}{\beta^2} \exp\left(-\frac{(e^v)^2}{2\beta^2}\right). \tag{A.2}$$

C'est ce que nous appelons la densité de probabilité d'une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2, ce que nous notons $V \sim \text{LogRay}(\beta^2)$. Cette densité de probabilité est tracée à la figure A.1. Donc, $Z = \ln Y$ est une variable aléatoire LogRayleigh circulaire de paramètre $\sigma^2/2$: $Z \sim \text{LogRay}(\sigma^2/2)$.

Propriété A.2 (Invariance des moments centrés supérieurs à 1)
Soit V une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2. Alors, tous les moments supérieurs centrés à 1 de V sont indépendants de β^2.

Preuve : [Propriété A.2]
Soient V_1 et V_2 deux variables LogRayleigh circulaires de paramètres de localisation respectifs β_1^2 et β_2^2, alors $p_{V_1}(\cdot)$ et $p_{V_2}(\cdot)$ vérifient

$$p_{V_2}(z + \zeta) = p_{V_1}(z), \quad \text{avec } \zeta = \ln\frac{\beta_2}{\beta_1}$$

Ceci signifie que chaque distribution se déduit des autres par une translation (*cf.* figure A.1). ■

De plus, soit V une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2. Sa moyenne m_V (l'unique moment centré qui dépend du paramètre de localisation β^2) de cette distribution est donnée par

$$m_V = \ln\beta + \frac{\ln 2}{2} - \frac{\gamma}{2} \tag{A.3}$$

où γ est la constante d'Euler définie par

$$\gamma = -\int_0^{+\infty} \ln x \, \exp(-x)\, dx,$$

son mode M_V est donné par

$$M_V = \ln \beta + \frac{\ln 2}{2} \tag{A.4}$$

et sa variance v_V vaut

$$v_V = \frac{\pi^2}{24}. \tag{A.5}$$

Preuve : Moyenne d'une variable aléatoire LogRayleigh circulaire
Par définition, on a

$$\begin{aligned} m_V &= E[v] = \int_{-\infty}^{+\infty} v\, p_c\left(v|\beta^2\right)\, dv \\ &= \int_{-\infty}^{+\infty} v\, \frac{(e^v)^2}{\beta^2}\, \exp\left(-\frac{(e^v)^2}{2\beta^2}\right)\, dv. \end{aligned}$$

Grâce au changement de variable

$$u^2 = \frac{(e^v)^2}{2\beta^2}$$

on obtient

$$\begin{aligned} m_V &= \int_0^{+\infty} \left[\ln(u) + \ln\left(\sqrt{2}\beta\right)\right]\, 2u \exp\left(-u^2\right)\, du \\ &= \ln\left(\sqrt{2}\beta\right) \underbrace{\int_0^{+\infty} 2u\, \exp\left(-u^2\right) du}_{1} + \int_0^{+\infty} 2u\, \ln(u)\, \exp\left(-u^2\right) du \\ &= \ln\beta + \frac{\ln 2}{2} + \frac{1}{2}\int_0^{+\infty} \ln(x)\, \exp(-x) dx \\ &= \ln\beta + \frac{\ln 2}{2} - \frac{\gamma}{2} \end{aligned}$$

■

Preuve : Mode d'une variable aléatoire LogRayleigh circulaire
Le mode est obtenu en annulant la dérivée de la densité de probabilité $p_V(\cdot)$:

$$\begin{aligned} \frac{dp_V}{dv}(y) &= \frac{1}{\beta^2}\, \exp\left(-\frac{e^{2y}}{2\beta^2}\right) \left[2\, e^{2v} - e^{2v}\, \frac{2\, e^{2v}}{2\beta^2}\right] \\ &= \frac{2\, e^{2v}}{\beta^2}\, \exp\left(-\frac{e^{2v}}{2\beta^2}\right) \left[1 - \frac{e^{2v}}{2\beta^2}\right] \end{aligned}$$

Donc, $M_v = \ln\beta + \ln 2/2$. ■

Preuve : Variance d'une variable aléatoire LogRayleigh circulaire
Par définition, on a

$$\begin{aligned} v_V &= \mathrm{Var}[v] = E[(v - m_V)^2] \\ &= E[v^2] - (m_V)^2 \\ &= \int_{-\infty}^{+\infty} v^2 \, \frac{(e^v)^2}{\beta^2} \, \exp\left(-\frac{(e^v)^2}{2\beta^2}\right) dv - \left[\ln\beta + \frac{\ln 2}{2} - \frac{\gamma}{2}\right]^2. \end{aligned}$$

Grâce au changement de variable

$$u^2 = \frac{(e^v)^2}{2\beta^2}$$

on obtient

$$\begin{aligned} v_V &= \int_0^{+\infty} \left[\ln u + \frac{\ln 2}{2} + \ln\beta\right]^2 \, 2u \exp\left(-u^2\right) du - \left[\ln\beta + \frac{\ln 2}{2} - \frac{\gamma}{2}\right]^2 \\ &= \int_0^{+\infty} (\ln u)^2 \, 2u \exp\left(-u^2\right) du + 2\left[\frac{\ln 2}{2} + \ln\beta\right] \int_0^{+\infty} \ln u \, 2u \exp\left(-u^2\right) du + \\ &\quad \left[\frac{\ln 2}{2} + \ln\beta\right]^2 \int_0^{+\infty} 2u \exp\left(-u^2\right) du - \left[\ln\beta + \frac{\ln 2}{2} - \frac{\gamma}{2}\right]^2 \\ &= \int_0^{+\infty} (\ln u)^2 \, 2u \exp\left(-u^2\right) du - \frac{\gamma^2}{4}. \end{aligned}$$

En utilisant la formule d'Euler-Mascheroni

$$\int_0^{+\infty} (\ln x)^2 \exp(-x)\, dx = \gamma^2 + \frac{\pi^4}{6}$$

on obtient

$$\begin{aligned} v_V &= \left(\frac{\pi^2}{24} + \frac{\gamma^2}{4}\right) - \frac{\gamma^2}{4} \\ &= \frac{\pi^2}{24}. \end{aligned}$$

■

Il est intéressant de noter que, puisque X est une variable aléatoire centrée complexe gaussienne circulaire, le carré de son module, Y^2, est relié à une distribution du chi-2 à deux degrés de liberté : $\chi^2(2)$ [92], qui est un cas particulier d'une distribution de gamma : $\Gamma(1, 1/2)$[92]. Ainsi, $\ln(Y^2)$, le logarithme de Y^2, est relié à la loi log-gamma [81]. Finalement, la distribution de Z, le logarithme du module de X, peut aussi être déduite de la distribution log-gamma par un changement de variable.

A.2 Conséquences de la non-circularité

A.2.1 Distribution de LogRayleigh non-circulaire

Dans ce paragraphe, nous dérivons la densité de probabilité d'une variable aléatoire V LogRayleigh non circulaire. Celle-ci est définie comme le logarithme du

module d'une variable aléatoire centrée complexe gaussienne non circulaire X. Les moments d'ordre deux de cette variable aléatoire X sont la covariance $v_X \triangleq \mathrm{E}[xx^*]$ (où * signifie le complexe conjugué) et la pseudo-covariance $c_X \triangleq \mathrm{E}[xx]$. Ainsi,

$$v_X = \sigma_r^2 + \sigma_i^2$$

et

$$c_X = \sigma_r^2 - \sigma_i^2 + 2\jmath\,\rho\,\sigma_r\,\sigma_i$$

où $\jmath = \sqrt{-1}$, ρ est le coefficient de corrélation entre les parties réelle et imaginaire de X, σ_r^2 et σ_i^2 sont respectivement les variances des parties réelle et imaginaire de X. Dans le cas circulaire, la pseudo-covariance est nulle ($c_X = 0$), ce qui signifie qu'à la fois $\sigma_{\Re(X)} = \sigma_{\Im(X)}$ et $\rho = 0$. Ainsi, la non-circularité peut se traduire par des covariances différentes entre les parties réelle et imaginaire et/ou à une corrélation entre celles-ci. Dans un souci de simplicité notons $\delta^2 = v_X$ et soit ϵ tel que

$$\sigma_r = \epsilon\ \sigma_i. \tag{A.6}$$

Dans ce cas, la densité de probabilité d'une variable aléatoire LR non circulaire Z est égale à

$$p_Z(z) = p_c\left(z\,\middle|\,\frac{\delta^2}{2}\right) I(z, \delta^2, \rho, \epsilon) \tag{A.7}$$

avec $p_c(z|\beta^2)$ la densité de probabilité d'une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2 donnée par l'équation (3.8) et

$$I(z, \delta^2, \rho, \epsilon) = \frac{\epsilon + 1/\epsilon}{2\sqrt{1-\rho^2}}\ \exp\left(-\frac{4\rho^2 + (\epsilon - 1/\epsilon)^2}{4(1-\rho^2)\delta^2}\,(e^z)^2\right) \times I_0\left(\left(\epsilon + \frac{1}{\epsilon}\right)\frac{\sqrt{(\epsilon - 1/\epsilon)^2 + 4\rho^2}}{4(1-\rho^2)\delta^2}(e^z)^2\right) \tag{A.8}$$

où $I_0(\cdot)$ est la fonction de Bessel modifiée de première espèce :

$$I_0(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp\{x\sin\theta\}d\theta.$$

Preuve :
Soient X_r et X_i les parties réelle et imaginaire de X une variable aléatoire centrée complexe gaussienne non circulaire. Dans ce cas, la densité de probabilité conjointe des parties réelle et imaginaire est donnée par

$$p_{X_r,X_i}(x_r, x_i) = \frac{1}{2\pi\,\sigma_r\,\sigma_i\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(\frac{x_r^2}{\sigma_r^2} - \frac{2\rho\,x_r\,x_i}{\sigma_r\,\sigma_i} + \frac{x_i^2}{\sigma_i^2}\right)\right].$$

Soient R et Θ le module et la phase de X. Nous avons alors

$$\begin{cases} x_r = r\ \cos\theta \\ x_i = r\ \sin\theta \end{cases}$$

ainsi

$$p_{R,\Theta}(r,\theta) = r\, p_{X_r,X_i}(r\cos\theta, r\sin\theta)$$

et

$$\begin{aligned}
p_R(r) &= \int_{-\pi}^{+\pi} r\, p_{X_r,X_i}(r\cos\theta, r\sin\theta)\, d\theta \\
&= \int_{-\pi}^{+\pi} r \frac{1}{2\pi\,\sigma_r\,\sigma_i\sqrt{1-\rho^2}} \times \\
&\quad \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(\frac{(r\,\cos\theta)^2}{\sigma_r^2} - \frac{2\rho\,(r\,\cos\theta)\,(r\,\sin\theta)}{\sigma_r\,\sigma_i} + \frac{(r\,\sin\theta)^2}{\sigma_i^2}\right)\right] d\theta \\
&= r\,h(r)\,\frac{\frac{\sigma_r}{\sigma_i}+\frac{\sigma_i}{\sigma_r}}{2\sqrt{1-\rho^2}}\,\exp\left[-r^2\frac{4\rho^2+\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2}{4(1-\rho^2)(\sigma_r^2+\sigma_i^2)}\right] \times \\
&\quad \frac{1}{2\pi}\int_{-\pi}^{+\pi} \exp\left[r^2\left(\frac{(\cos^2\theta-\sin^2\theta)\left(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_r^2}\right)}{4(1-\rho^2)} + \frac{\rho\sin\theta\cos\theta}{(1-\rho^2)\sigma_r\sigma_i}\right)\right] d\theta
\end{aligned}$$

où $h(\cdot)$ est donné par

$$h(r) = \frac{r}{\frac{\sigma_r^2+\sigma_i^2}{2}}\,\exp\left[-\frac{r^2}{2\frac{\sigma_r^2+\sigma_i^2}{2}}\right]$$

Soit $Z = \ln R$ la variable aléatoire LogRayleigh non-circulaire, on a alors

$$\begin{aligned}
p_Z(z) &= p_c\left(z \left| \frac{\sigma_r^2+\sigma_i^2}{2}\right.\right) \frac{\frac{\sigma_r}{\sigma_i}+\frac{\sigma_i}{\sigma_r}}{2\sqrt{1-\rho^2}} \exp\left[-(e^z)^2 \frac{4\rho^2+\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2}{4(1-\rho^2)(\sigma_r^2+\sigma_i^2)}\right] \times \\
&\quad \frac{1}{2\pi}\int_{-\pi}^{+\pi} \exp\left[(e^z)^2\left(\frac{\cos 2\theta\left(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_r^2}\right)}{4(1-\rho^2)} + \frac{\rho \sin 2\theta}{2(1-\rho^2)\sigma_r\sigma_i}\right)\right] d\theta \\
&= p_c\left(z \left| \frac{\sigma_r^2+\sigma_i^2}{2}\right.\right) \frac{\frac{\sigma_r}{\sigma_i}+\frac{\sigma_i}{\sigma_r}}{2\sqrt{1-\rho^2}} \exp\left[-(e^z)^2 \frac{4\rho^2+\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2}{4(1-\rho^2)(\sigma_r^2+\sigma_i^2)}\right] \times \\
&\quad \frac{1}{2\pi}\int_{-\pi}^{+\pi} \exp\left[(e^z)^2 \sqrt{\frac{\left(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_r^2}\right)^2}{16\left(1-\rho^2\right)^2} + \frac{\rho^2}{4\left(1-\rho^2\right)^2\sigma_r^2\sigma_i^2}} \sin 2\theta\right] d\theta \\
&= p_c\left(z \left| \frac{\sigma_r^2+\sigma_i^2}{2}\right.\right) \frac{\frac{\sigma_r}{\sigma_i}+\frac{\sigma_i}{\sigma_r}}{2\sqrt{1-\rho^2}} \exp\left[-(e^z)^2 \frac{4\rho^2+\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2}{4(1-\rho^2)(\sigma_r^2+\sigma_i^2)}\right] \times \\
&\quad \frac{1}{2\pi}\int_{-\pi}^{+\pi} \exp\left[(e^z)^2 \sqrt{\frac{\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2+4\rho^2}{4\left(1-\rho^2\right)\sigma_r\sigma_i}} \sin\theta\right] d\theta \\
&= p_c\left(z \left| \frac{\sigma_r^2+\sigma_i^2}{2}\right.\right) \frac{\frac{\sigma_r}{\sigma_i}+\frac{\sigma_i}{\sigma_r}}{2\sqrt{1-\rho^2}} \exp\left[-(e^z)^2 \frac{4\rho^2+\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2}{4(1-\rho^2)(\sigma_r^2+\sigma_i^2)}\right] \times \\
&\quad I_0\left((e^z)^2 \sqrt{\frac{\left(\frac{\sigma_r}{\sigma_i}-\frac{\sigma_i}{\sigma_r}\right)^2+4\rho^2}{4\left(1-\rho^2\right)\sigma_r\sigma_i}}\right)
\end{aligned}$$

On a alors en substituant $\sigma_r^2+\sigma_i^2$ par δ^2 et σ_r/σ_i par ϵ

$$p_Z(z) = p_c\left(z \left| \frac{\delta^2}{2}\right.\right) I(z, \delta^2, \rho, \epsilon) \tag{A.9}$$

avec $p_c(z|\beta^2)$ la densité de probabilité d'une variable aléatoire LogRayleigh circulaire de paramètre de localisation β^2 donnée par l'équation (3.8) et

$$\begin{aligned}
I(z, \delta^2, \rho, \epsilon) = \frac{\epsilon + 1/\epsilon}{2\sqrt{1-\rho^2}} &\exp\left(-\frac{4\rho^2+(\epsilon-1/\epsilon)^2}{4(1-\rho^2)\delta^2}(e^z)^2\right) \\
&\times I_0\left(\left(\epsilon+\frac{1}{\epsilon}\right)\frac{\sqrt{(\epsilon-1/\epsilon)^2+4\rho^2}}{4(1-\rho^2)\delta^2}(e^z)^2\right).
\end{aligned} \tag{A.10}$$

■

A.2.2 Calcul du paramètre de localisation optimal

Dans ce paragraphe, nous calculons le paramètre de localisation optimal entre une distribution LogRayleigh non-circulaire $p_Z(\cdot)$ et la distribution LogRayleigh circulaire $p_c(\cdot|\alpha)$ de paramètre de localisation optimal α au sens de la divergence de Kullback-Leibler :

$$\hat{\alpha} = \arg\min_{\alpha} g(\alpha)$$

où

$$g(\alpha) = KL[p_Z(\cdot)\|p_c(\cdot|\alpha)] = \int_{-\infty}^{+\infty} p_Z(z)\ \ln\left(\frac{p_Z(z)}{p_c(z|\alpha)}\right) dz.$$

On a alors

$$\hat{\alpha} = \frac{\delta^2}{2}. \tag{A.11}$$

Preuve :
Le paramètre optimal $\hat{\alpha}$ annule le gradient de $g(\alpha)$. Or, on a

$$\begin{aligned}\frac{\partial g(\alpha)}{\partial \alpha} &= -\int_{-\infty}^{+\infty} p_Z(z)\,\frac{\partial}{\partial \alpha}\left[\ln p_c(z|\alpha)\right] dz \\ &= -\int_{-\infty}^{+\infty} p_Z(z)\left[-\frac{1}{\alpha} + \frac{(e^z)^2}{2\alpha^2}\right] dz\end{aligned}$$

ainsi

$$\hat{\alpha} = \frac{1}{2}\int_{-\infty}^{+\infty} (e^z)^2\, p_Z(z) dz. \tag{A.12}$$

En substituant $p_Z(z)$ par l'équation (A.9), on obtient

$$\begin{aligned}\hat{\alpha} &= \frac{1}{2}\int_{-\infty}^{+\infty} (e^z)^2\, p_c\left(z\left|\frac{\delta^2}{2}\right.\right) I(z,\delta^2,\rho,\epsilon)\, dz \\ &= \frac{1}{2}\int_{-\infty}^{+\infty} (e^z)^2\, p_c\left(z\left|\frac{\delta^2}{2}\right.\right) \frac{\epsilon + 1/\epsilon}{2\sqrt{1-\rho^2}}\ \exp\!\left(-\frac{4\rho^2 + (\epsilon - 1/\epsilon)^2}{4(1-\rho^2)\delta^2}\,(e^z)^2\right) \times \\ &\quad I_0\left(\left(\epsilon + \frac{1}{\epsilon}\right)\frac{\sqrt{(\epsilon - 1/\epsilon)^2 + 4\rho^2}}{4(1-\rho^2)\delta^2}\ (e^z)^2\right).\end{aligned}$$

Par le changement de variable $u = (e^z)^2$, on aboutit à

$$\begin{aligned}\hat{\alpha} &= \frac{1}{2}\int_{u=0}^{u=+\infty} \frac{u}{\delta^2}\ \exp\!\left[-\frac{u}{\delta^2}\right] \frac{\epsilon + 1/\epsilon}{2\sqrt{1-\rho^2}}\ \exp\!\left[-u\frac{4\rho^2 + (\epsilon - 1/\epsilon)^2}{4(1-\rho^2)\delta^2}\right] \times \\ &\quad \frac{1}{2\pi}\int_{\theta=-\pi}^{\theta=+\pi} \exp\left[u\frac{\sqrt{(\epsilon - 1/\epsilon)^2 + 4\rho^2}}{4(1-\rho^2)\delta^2}\left(\frac{1}{\epsilon} + \epsilon\right)\sin\theta\right] d\theta\, du\end{aligned}$$

d'où, après intégration sur u

$$\hat{\alpha} = \frac{\delta^2}{2}\, J(\rho,\epsilon)$$

avec

$$J(\rho,\epsilon) = \left(\epsilon + \frac{1}{\epsilon}\right)\left(1-\rho^2\right)^{3/2}\ \frac{4}{\pi}\ \int_{-\pi}^{\pi} \frac{d\theta}{(a - b\,\sin\theta)^2} \tag{A.13}$$

où

$$a = 4 + \left(\epsilon - \frac{1}{\epsilon}\right)^2 = \left(\epsilon + \frac{1}{\epsilon}\right)^2 \tag{A.14}$$

et

$$b = \left(\epsilon + \frac{1}{\epsilon}\right)\sqrt{\left(\epsilon - \frac{1}{\epsilon}\right)^2 + 4\rho^2}. \tag{A.15}$$

Grâce à la formule (14.361) de [122], on peut montrer que

$$\int_{-\pi}^{\pi} \frac{d\theta}{(a - b\,\sin\theta)^2} = \frac{a}{(a^2 - b^2)} \int_{-\pi}^{\pi} \frac{d\theta}{a - b\,\sin\theta} \tag{A.16}$$

et grâce à la formule (14.360) de [122]

$$\int_{-\pi}^{\pi} \frac{d\theta}{a - b\,\sin\theta} = \frac{2\,\pi}{\sqrt{a^2 - b^2}}. \tag{A.17}$$

Ainsi, en substituant les expressions (A.16) et (A.17) dans l'équation (A.13), on obtient

$$J(\epsilon,\rho) \;=\; \left(\epsilon + \frac{1}{\epsilon}\right)\left(1 - \rho^2\right)^{3/2} \frac{8a}{(a^2 - b^2)^{3/2}}.$$

En remarquant que

$$\begin{aligned}
(a^2 - b^2) &= (a - b)(a + b) \\
&= \left(\epsilon + \frac{1}{\epsilon}\right)\left[\left(\epsilon + \frac{1}{\epsilon}\right) + \sqrt{\left(\epsilon - \frac{1}{\epsilon}\right)^2 + 4\rho^2}\right] \times \\
&\quad \left(\epsilon + \frac{1}{\epsilon}\right)\left[\left(\epsilon + \frac{1}{\epsilon}\right) - \sqrt{\left(\epsilon - \frac{1}{\epsilon}\right)^2 + 4\rho^2}\right] \\
&= \left(\epsilon + \frac{1}{\epsilon}\right)^2\left[\left(\epsilon + \frac{1}{\epsilon}\right)^2 - \left(\epsilon - \frac{1}{\epsilon}\right)^2 - 4\rho^2\right] \\
&= 4\left(\epsilon + \frac{1}{\epsilon}\right)^2\left(1 - \rho^2\right).
\end{aligned}$$

On aboutit ainsi à

$$\begin{aligned}
J(\epsilon,\rho) &= \left(\epsilon + \frac{1}{\epsilon}\right)\left(1 - \rho^2\right)^{3/2} \frac{8\left(\epsilon + \frac{1}{\epsilon}\right)^2}{\left(4\left(\epsilon + \frac{1}{\epsilon}\right)^2\left(1 - \rho^2\right)\right)^{3/2}} \\
&= 1.
\end{aligned}$$

Finalement on obtient

$$\hat{\alpha} = \frac{\delta^2}{2}. \tag{A.18}$$

■

A.3 Conditionnement numérique des paramètres

Comparons dans ce paragraphe le conditionnement numérique du logarithme des modules des coefficients spectraux vis-à-vis de ces mêmes coefficients spectraux. Soit $\mathbf{S}(t)$ le vecteur des coefficients complexes de la TFCT d'un signal $s(t)$. Pour des sections de parole quasi-stationnaires, les coefficients complexes $\mathbf{S}(t) = [S(t,f_1),\cdots,S(t,f_{N_f})]^T$ de la TFD peuvent être considérés comme décorrélés et ayant une distribution complexe gaussienne circulaire centrée [89, 103] (certains auteurs [89] préfèrent employer le terme propre, *proper* en anglais, à la place de circulaire) :

$$p\big(\mathbf{S}(t)\big) = p_G\big(\mathbf{S}(t)|0,\Sigma^A\big) = \prod_{j=1}^{N_f} p_G\big(S(t,f_j)\big|0,\Sigma^A(f_j)\big). \tag{A.19}$$

Soit $\mathbf{a}(t) = \ln\big|\mathbf{S}(t)\big|$ le vecteur du logarithme du module de $\mathbf{S}(t)$. Ce vecteur suit donc une loi LogRayleigh circulaire (*cf.* paragraphe 3.3.1 page 60)

$$p\big(\mathbf{a}(t)\big) = p_{LR}\big(\mathbf{a}(t)\big|\Gamma^A\big) = \prod_{j=1}^{N_f} p_{LR}\big(a(t,f_j)\,|\Gamma^A(f_j)\big)\,, \tag{A.20}$$

où $\Gamma^A(f) = \Sigma^A(f)/2$ et $\Sigma^A(f)$ est la variance de $S(t,f)$.

Soit $s'(t)$ le même son mais de puissance différente : $s'(t) = \alpha\, s(t)$. Les coefficients de la TFCT de $s'(t)$ sont donc donnés par $\mathbf{S}'(t) = \alpha\,\mathbf{S}(t)$ et suivent une loi complexe gaussienne circulaire centrée telle que

$$p\big(\mathbf{S}'(t)\big) = p_G\big(\mathbf{S}'(t)\big|0,\alpha^2\,\Sigma^A\big) = \prod_{j=1}^{N_f} p_G\big(S'(t,f_j)\big|0,\alpha^2\,\Sigma^A(f_j)\big). \tag{A.21}$$

Ainsi, la loi du logarithme du module de ces coefficients $\mathbf{a}'(t) = \ln\big|\mathbf{S}'(t)\big|$ vérifie

$$p\big(\mathbf{a}'(t)\big) = p_{LR}\big(\mathbf{a}(t)\big|\alpha^2\,\Gamma^A\big) = \prod_{j=1}^{N_f} p_{LR}\big(a(t,f_j)\,|\alpha^2\,\Gamma^A(f_j)\big)\,. \tag{A.22}$$

On peut montrer d'après (A.19) et (A.21) que

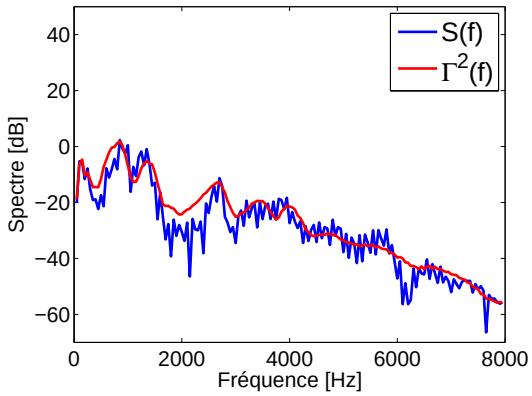
$$p\big(\mathbf{S}'(t)\big) = \left(\frac{1}{\alpha^2}\right)^{N_f} p\big(\mathbf{S}(t)\big).$$

Donc pour un même son mais à des puissances différentes, la valeur prise par la densité de probabilité modélisant directement les coefficients complexes de la TFCT $\mathbf{S}(t)$ varie en fonction de la puissance du signal (*cf.* figure A.2). En d'autres termes, changer la puissance du signal change la valeur de la vraisemblance des coefficients de la TFCT. En revanche, la propriété d'invariance des moments d'ordre supérieur à deux d'une variable aléatoire LogRayleigh, démontrée au paragraphe A.1, permet de s'affranchir de ce problème. En effet, on peut montrer que

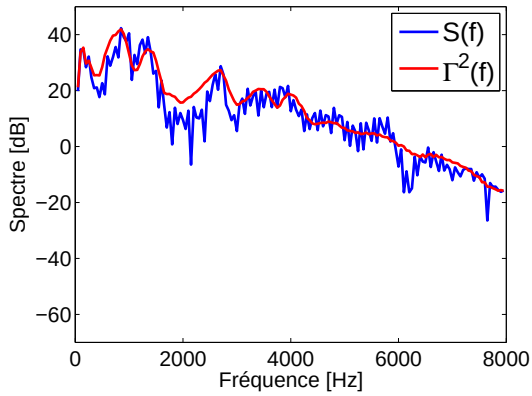
$$p\big(\mathbf{a}'(t)\big) = p\big(\mathbf{a}(t)\big)$$

$\mathcal{P}$ [dB]	-20	-5	0	5	20
G (A.21)	$1.97\ 10^{292}$	$6.2425\ 10^{53}$	$1.97\ 10^{-26}$	$6.2425\ 10^{-106}$	0
LR (A.22)	$5.86\ 10^{-62}$	$5.86\ 10^{-62}$	$5.86\ 10^{-62}$	$5.86\ 10^{-62}$	$5.86\ 10^{-62}$

(a)



(b) $\mathcal{P} = -20$dB



(c) $\mathcal{P} = 20$dB

FIG. A.2 – Conditionnement numérique. Le tableau A.2(a) regroupe les valeurs prises par les vraisemblances gaussiennes (A.21) et LogRayleigh (A.22) pour différentes valeurs de la puissance moyenne $\mathcal{P}$ d'un même signal. Les figures A.2(b) et A.2(b) montre, pour deux valeurs de la puissance moyenne $\mathcal{P}$, le logarithme des coefficients de la TFCT $\ln\left|\mathbf{S}(t)\right|$ (tracé bleu) et le paramètre de localisation Γ^A (tracé rouge) en fonction de la fréquence.

puisque les distributions de deux variables LR de paramètre de localisation différent (et donc de puissance moyenne différente) se déduisent l'une de l'autre par translation. En d'autres termes, changer la puissance du signal laisse inchangée la valeur de la vraisemblance du logarithme du module des coefficients de la TFCT. Ceci est illustré à la figure A.2.

Annexe B

Algorithme EM

Dans cette annexe, nous décrivons l'algorithme EM [46] dans sa forme pénalisée [91, 116] pour l'estimation des paramètres du modèle audiovisuel proposé au paragraphe 3.3 de la partie II sur la modélisation de la bimodalité de la parole.

B.1 Principe de l'algorithme EM

Soit X une variable aléatoire régie par une loi telle que

$$X \sim \sum_{i=1}^{I} \omega_i \; p(\mathbf{x}|\theta_i) \qquad \text{avec} \qquad \sum_{i=1}^{I} \omega_i = 1 \tag{B.1}$$

où $p(\mathbf{x}|\theta_i)$ est une loi quelconque dépendant du paramètre θ_i. ω_i est le poids de la loi indexée i. On a ainsi

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{I} \omega_i \; p(\mathbf{x}|\theta_i) \tag{B.2}$$

où $\Theta = \{\omega_i, \theta_i\}_{1 \leq i \leq I}$.

L'algorithme EM [46] permet d'estimer de façon itérative le paramètre Θ du modèle (B.1) par la méthode du maximum de vraisemblance et assure la convergence vers un maximum local.

Pour cela, supposons que nous ayons à notre disposition un ensemble $\mathcal{X}$ de T données générées par la distribution (B.1) : $\mathcal{X} = \{\mathbf{x}(1), \cdots, \mathbf{x}(T)\}$. En faisant l'hypothèse que ces données sont indépendantes entre elles, nous pouvons écrire

$$p(\mathcal{X}|\Theta) = \prod_{t=1}^{T} p(\mathbf{x}(t)|\Theta) = \mathcal{L}(\Theta|\mathcal{X}) \tag{B.3}$$

où $\mathcal{L}(\Theta|\mathcal{X})$ est la vraisemblance du paramètre Θ conditionnée aux données $\mathcal{X}$.

Dans le problème du maximum de vraisemblance, un estimateur $\hat{\Theta}$ du paramètre Θ est donné par la maximisation de $\mathcal{L}(\Theta|\mathcal{X})$:

$$\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\Theta|\mathcal{X}) \tag{B.4}$$

Dans certains cas où la fonction de vraisemblance $\mathcal{L}(\Theta|\mathcal{X})$ est complexe[1], l'algorithme EM permet d'estimer le paramètre Θ en supposant l'existence de paramètres cachés inconnus. Ainsi, soient $\mathcal{X}$ l'ensemble des données observées mais incomplètes et $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ les données complètes où $\mathcal{Y}$ représente les paramètres cachés inconnus. Nous avons alors

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, y|\Theta). \tag{B.5}$$

B.1.1 Algorithme EM standard

L'algorithme EM [46] procède en deux étapes itérées autant que fois que nécessaires :

1. l'étape (E) consiste à estimer la log-vraisemblance des données complètes sachant les données observées $\mathcal{X}$ et le paramètre courant $\Theta^{(k)}$ obtenu à l'étape k

$$Q(\Theta, \Theta^{(k)}) = E\left\{\ln\left[p(\mathcal{X}, \mathcal{Y}|\Theta)\right] \middle| \mathcal{X}, \Theta^{(k)}\right\} \tag{B.6}$$

2. l'étape (M) consiste à maximiser l'espérance calculée à l'étape (E)

$$\Theta^{(k+1)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(k)}). \tag{B.7}$$

En utilisant le modèle (B.1), la log-vraisemblance des données incomplètes s'écrit

$$\begin{aligned} \ln\left[\mathcal{L}(\Theta|\mathcal{X})\right] &= \ln\left[\prod_{t=1}^{T} p(\mathbf{x}(t)|\Theta)\right] \\ &= \sum_{t=1}^{T} \ln\left[\sum_{i=1}^{I} \omega_i\, p(\mathbf{x}(t)|\theta_i)\right] \end{aligned}$$

expression qu'il n'est pas aisée de maximiser par rapport à Θ. En revanche, en considérant les données complètes $\mathcal{Z}$ où la donnée cachée $y(t)$ représente le noyau de la densité (B.1) qui a généré la donnée $\mathbf{x}(t)$, la log-vraisemblance devient

$$\begin{aligned} \ln\left[\mathcal{L}(\Theta|\mathcal{Z})\right] &= \ln\left[\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})\right] \\ &= \ln\left[\prod_{t=1}^{T} p(\mathbf{x}(t), y(t)|\Theta)\right] \\ &= \sum_{t=1}^{T} \ln\left[p(y(t)|\Theta)\, p(\mathbf{x}(t)|y(t), \Theta)\right] \end{aligned}$$

ce qui donne donc

$$\ln\left[\mathcal{L}(\Theta|\mathcal{Z})\right] = \sum_{t=1}^{T} \ln\left[\omega_{y(t)}\, p(\mathbf{x}(t)|\theta_{y(t)})\right], \tag{B.8}$$

[1] C'est le cas notamment de densités de probabilité définies comme mélange de densités élémentaires (par exemple le cas des mélanges de gaussiennes).

expression qu'il est plus facile de maximiser par rapport aux paramètres si l'on connaît les données cachées $\mathcal{Y} = \mathbf{y} = \{y(t), \cdots, y(T)\}$.

De façon à surmonter cette difficulté, supposons que $\mathbf{y}$ soit un vecteur aléatoire. La règle de Bayes donne ainsi

$$p(y(t)|\mathbf{x}(t), \Theta) = \frac{p(\mathbf{x}(t)|y(t), \Theta)\, p(y(t)|\Theta)}{p(\mathbf{x}(t)|\Theta)}$$

que l'on peut aussi écrire

$$p(y(t)|\mathbf{x}(t), \Theta) = \frac{p(\mathbf{x}(t)|\theta_{y(t)})\ \omega_{y(t)}}{\sum_{i=1}^{I} \omega_i\, p(\mathbf{x}(t)|\theta_i)}. \tag{B.9}$$

De plus on a

$$p(\mathbf{y}|\mathcal{X}, \Theta) = \prod_{t=1}^{T} p(y(t)|\mathbf{x}(t), \Theta). \tag{B.10}$$

Or, l'équation (B.6) de l'étape (E) de l'algorithme EM s'écrit :

$$\begin{aligned} Q(\Theta, \Theta^{(k)}) &= E\left\{\ln\left[p(\mathcal{X}, \mathbf{y}|\Theta)\right] \middle| \mathcal{X}, \Theta^{(k)}\right\} \\ &= E\left\{\ln\left[\prod_{t=1}^{T} \left(p(y(t)|\Theta)\ p(\mathbf{x}(t)|y(t), \Theta)\right)\right] \middle| \mathcal{X}, \Theta^{(k)}\right\} \\ &= E\left\{\sum_{t=1}^{T} \ln\left[\omega_{y(t)}\, p(\mathbf{x}(t)|\theta_{y(t)})\right] \middle| \mathcal{X}, \Theta^{(k)}\right\} \\ &= \sum_{t=1}^{T}\sum_{i=1}^{I} \ln\left[\omega_i\, p(\mathbf{x}(t)|\theta_i)\right]\ p\left(i \middle| \mathbf{x}(t), \Theta^{(k)}\right). \end{aligned}$$

On a donc

$$Q(\Theta, \Theta^{(k)}) = \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{I} \ln\left[\omega_i\right]\ p\left(i \middle| \mathbf{x}(t), \Theta^{(k)}\right)}_{\Phi(\omega)} + \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{I} \ln\left[p(\mathbf{x}(t)|\theta_i)\right]\ p\left(i \middle| \mathbf{x}(t), \Theta^{(k)}\right)}_{\Psi(\theta)}$$

où $\omega = [\omega_1, \cdots, \omega_I]^T$ et $\theta = [\theta_1, \cdots, \theta_I]^T$. On constate alors que maximiser cette expression par rapport à Θ pour $\Theta^{(k)}$ fixé est équivalent à maximiser $\Phi(\cdot)$ par rapport à l'ensemble des ω_i et à maximiser $\Psi(\cdot)$ par rapport à l'ensemble des θ_i séparemment puisqu'ils sont disjoints.

Pour estimer les ω_i, nous introduisons un terme de contrainte λ afin d'assurer que $\sum_{i=1}^{I} \omega_i = 1$. On a ainsi pour tout j :

$$\frac{\partial}{\partial \omega_j}\left[\Phi(\omega) + \lambda\left(\sum_{i=1}^{I} \omega_i - 1\right)\right] = 0$$

or

$$\frac{\partial}{\partial \omega_j}\left[\Phi(\omega) + \lambda\left(\sum_{i=1}^{I} \omega_i - 1\right)\right] = \frac{\partial}{\partial \omega_j}\left[\sum_{t=1}^{T}\sum_{i=1}^{I} \ln[\omega_i]\ p\big(i|\mathbf{x}(t), \Theta^{(k)}\big) + \lambda\left(\sum_{i=1}^{I} \omega_i - 1\right)\right]$$

ce qui donne

$$\forall j, \qquad \frac{1}{\omega_j} \sum_{t=1}^{T} p\big(j|\mathbf{x}(t), \Theta^{(k)}\big) + \lambda = 0 \tag{B.11}$$

en sommant toutes les équations (B.11) sur j, nous obtenons que $\lambda = -T$, ce qui fournit alors

$$\forall j, \qquad \omega_j = \frac{1}{T} \sum_{t=1}^{T} p(j|\mathbf{x}(t), \Theta^{(k)}) \tag{B.12}$$

L'expression des θ_i sera estimée par l'annulation de la dérivé de $\Psi(\theta)$ par rapport aux θ_j :

$$\forall j, \qquad \frac{\partial \Psi(\theta)}{\partial \theta_j} = 0 \tag{B.13}$$

Cette expression dépend des lois élémentaires $p(\mathbf{x}(t)|\theta_i)$ et une expression analytique existe dans certains cas.

B.1.2 Algorithme EM pénalisé

Si la vraisemblance des paramètres Θ n'est pas bornée[2] alors des problèmes numériques peuvent survenir. Ceci est bien connu et une manière de lever cette dégénérescence consiste à ajouter une loi *a priori* sur les paramètres : on utilise alors l'algorithme EM pénalisé [91, 116].

Soit $p(\Theta)$ la loi *a priori* des paramètres. Le principe de l'algorithme EM pénalisé consiste à estimer les paramètres Θ à partir de la vraisemblance pénalisée :

$$\hat{\Theta} = \arg\max_{\Theta} \ln \left[p(\mathcal{X}, \mathcal{Y}|\Theta)\ p(\Theta)\right]. \tag{B.14}$$

Il s'agit alors d'un estimateur équivalent au maximum *a posteriori*. On utilise, tout comme pour l'algorithme EM standard, deux étapes :

1. l'étape (E) qui consiste à estimer la log-vraisemblance pénalisée des données complètes sachant les données observées $\mathcal{X}$ et le paramètre courant $\Theta^{(k)}$

$$Q(\Theta, \Theta^{(k)}) = E\left\{\ln\left[p(\mathcal{X}, \mathcal{Y}|\Theta)\ p(\Theta)\right] \big| \mathcal{X}, \Theta^{(k)}\right\} \tag{B.15}$$

2. l'étape (M) qui consiste à maximiser l'espérance calculée à l'étape (E)

$$\Theta^{(k+1)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(k)}). \tag{B.16}$$

On a alors

$$\begin{aligned} Q(\Theta, \Theta^{(k)}) &= E\left\{\ln\left[p(\mathcal{X}, \mathbf{y}|\Theta)\ p(\Theta)\right] \big| \mathcal{X}, \Theta^{(k)}\right\} \\ &= \sum_{t=1}^{T} \sum_{i=1}^{I} \ln\left[\omega_i\, p(\mathbf{x}(t)|\theta_i)\right]\ p\big(i\,\big|\mathbf{x}(t), \Theta^{(k)}\big) + \ln[p(\Theta)] \end{aligned}$$

[2] C'est notamment ce qui arrive si les lois élémentaires du modèle (B.1) sont des gaussiennes et qu'au moins une variance tend vers 0.

on a donc

$$\begin{aligned} Q(\Theta,\Theta^{(k)}) &= \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{I}\ln\left[\omega_i\right]\; p\left(i\,\middle|\mathbf{x}(t),\Theta^{(k)}\right)+\ln[p(\omega)]}_{\Phi(\omega)}+\cdots \\ &+ \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{I}\ln\left[p(\mathbf{x}(t)|\theta_i)\right]\; p\left(i\,\middle|\mathbf{x}(t),\Theta^{(k)}\right)+\ln[p(\theta)]}_{\Psi(\theta)} \end{aligned} \tag{B.17}$$

si l'on choisit une loi *a priori* séparable : $p(\Theta)=p(\omega)\,p(\theta)$. De plus, adopter pour loi *a priori* la loi conjuguée de la vraisemblance permet dans certain cas d'obtenir des expressions analytiques des nouveaux paramètres $\Theta^{(k+1)}$ à l'itération $k+1$.

B.2 Algorithme EM pour le modèle audiovisuel

Nous allons maintenant appliquer le principe de l'algorithme EM pénalisé dans le cadre du modèle audiovisuel défini par l'équation (3.18) que nous rappelons ici

$$p_{AV}(\mathbf{a}(t),\mathbf{v}(t))=\sum_{i=1}^{N_{AV}}\omega_i^{AV}\,p_G\big(\mathbf{v}(t)\,\big|\mu_i^V,\Sigma_i^V\big)\;p_{LR}\big(\mathbf{a}(t)\,\big|\Gamma_i^A\big)\,. \tag{B.18}$$

Dans ce cas particulier, l'ensemble Θ des paramètres à estimer est donné par

$$\Theta=\{\omega_i^{AV},\mu_i^V,\Sigma_i^V,\Gamma_i^A\}_{1\leq i\leq N_{AV}}$$

où le nombre de noyaux audiovisuel N_{AV} est fixé *a priori*, Σ_i^V sont les matrices de covariance vidéo et les matrices de localisation audio sont également diagonales $\Gamma_i^A=\mathrm{diag}(\Gamma_i^A(f_1),\cdots,\Gamma_i^A(f_{N_f}))$. Le fait que les noyaux audiovisuels soient séparables entraîne que $\Psi(\theta)$ défini à l'équation (B.17), avec $\theta=\{\mu_i^V,\Sigma_i^V,\Gamma_i^A\}_i$, puisse faire intervenir les paramètres vidéo et audio séparément. En effet,

$$\begin{aligned} \Psi(\theta) &= \sum_{t=1}^{T}\sum_{i=1}^{N_{AV}}\left[\ln\left[p(\mathbf{a}(t),\mathbf{v}(t)|\theta_i)\right]\,p\big(i\,\big|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)}\big)\right]+\ln[p(\theta)] \\ &= \sum_{t=1}^{T}\sum_{i=1}^{N_{AV}}\left[p_G(\mathbf{v}(t)|\mu_i^V,\Sigma_i^V)\,p\big(i\,\big|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)}\big)\right]+\sum_{i=1}^{N_{AV}}\ln[p(\mu_i^V,\Sigma_i^V)]+ \\ &\quad \sum_{t=1}^{T}\sum_{i=1}^{N_{AV}}\left[p_{LR}\big(\mathbf{a}(t)\,\big|\Gamma_i^A\big)\;p\big(i\,\big|\mathbf{a}(t),\mathbf{v}(t),\Theta^{(k)}\big)\right]+\sum_{i=1}^{N_{AV}}\ln[p(\Gamma_i^A)] \end{aligned} \tag{B.19}$$

si l'on choisit des densités *a priori* totalement séparables pour θ. Ceci montre que l'on peut mettre à jour les paramètres vidéo et audio séparément.

B.2.1 Mise à jour des poids

L'équation qui permet de mettre à jour les poids à l'itération $k+1$ à partir du paramètre $\Theta^{(k)}$ estimé à l'itération précédente k est donnée par la fonction $\Phi(\omega^{AV})$ définie à l'équation (B.17), où $\omega^{AV} = [\omega_1^{AV}, \cdots, \omega_{N_{AV}}^{AV}]^T$:

$$\Phi\big(\omega^{AV}\big) = \sum_{t=1}^{T}\sum_{i=1}^{N_{AV}} \ln\big[\omega_i^{AV}\big]\ \ p\big(i\,\big|\mathbf{x}(t), \Theta^{(k)}\big) + \ln[p(\omega^{AV})].$$

La distribution conjuguée des poids ω_i^{AV} est une distribution de Dirichlet [91] telle que

$$D\big(\omega^{AV}\,|\kappa\big) \propto \prod_{i=1}^{N_{AV}} \big(\omega_i^{AV}\big)^{\kappa_i - 1}.$$

On a alors

$$\Phi\big(\omega^{AV}\big) = \sum_{t=1}^{T}\sum_{i=1}^{N_{AV}} \ln\big[\omega_i^{AV}\big]\ \ p\big(i\,\big|\mathbf{x}(t), \Theta^{(k)}\big) + \ln\big[D\big(\omega^{AV}\,|\kappa\big)\big].$$

Pour assurer que $\sum_{i=1}^{N_{AV}} \omega_i^{AV} = 1$, il est possible d'introduire un facteur de contrainte λ, on a alors

$$\begin{aligned}
\frac{\partial}{\partial \omega_j^{AV}}\left[\Phi\big(\omega^{AV}\big) + \lambda\left(\sum_{i=1}^{N_{AV}} \omega_i^{AV} - 1\right)\right] &= 0 \\
\frac{\partial}{\partial \lambda}\left[\Phi\big(\omega^{AV}\big) + \lambda\left(\sum_{i=1}^{N_{AV}} \omega_i^{AV} - 1\right)\right] &= 0.
\end{aligned}$$

d'où

$$\frac{1}{\omega_j^{AV}}\sum_{t=1}^{T} p\big(i\,\big|\mathbf{x}(t), \Theta^{(k)}\big) + \frac{1}{\omega_j^{AV}}\,(\kappa_j - 1) + \lambda = 0$$

avec

$$\sum_{i=1}^{N_{AV}} \omega_i^{AV} - 1 = 0.$$

Finalement l'équation de mise à jour des poids est donnée par

$$\big(\omega_i^{AV}\big)^{(k+1)} = \frac{\sum_{t=1}^{T} p\big(i\,\big|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big) + \kappa_i - 1}{T + \sum_{j=1}^{N_{AV}} \kappa_j - N_{AV}}. \tag{B.20}$$

Sans information *a priori* sur les fréquences d'apparition des sons dans le corpus, nous adoptons un *a priori* non informatif sur la répartition de ceux-ci : $\kappa_i = 1$ pour tout i. Ainsi, on obtient la même expression que dans le cas standard

$$\big(\omega_i^{AV}\big)^{(k+1)} = \frac{1}{T}\sum_{t=1}^{T} p\big(i\,\big|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big). \tag{B.21}$$

B.2.2 Mise à jour des paramètres vidéo

Les équations qui permettent de mettre à jour les paramètres vidéo sont obtenues en annulant les dérivées de (B.19) par rapport à μ_i^V et Σ_i^V respectivement. Les distributions conjuguées sur les vecteurs des valeurs moyennes μ_i^V et les matrices de covariance Σ_i^V sont respectivement une distribution normale et une distribution de Wishart inverse [91, 116], on a alors

$$\begin{aligned} \mu_i^V &\sim \mathcal{N}\big(\nu_i, \eta_i^{-1}\Sigma_i^V\big) \\ \Sigma_i^V &\sim \mathcal{W}i(\alpha_i, \beta_i, J_i) \end{aligned}$$

où ν_i et η_i sont respectivement un vecteur et un scalaire, α_i et β_i sont des scalaires tandis que les J_i sont des matrices symétriques positives. Ces distributions sont telles que

$$p\big(\mu_i^V \,\big|\nu_i, \eta_i^{-1}\Sigma_i^V\big) \propto \frac{1}{\sqrt{\det\left(\eta_i^{-1}\Sigma_i^V\right)}} \exp\left[-\frac{\eta_i}{2}\left(\mu_i^V - \nu_i\right)^T \left(\Sigma_i^V\right)^{-1}\left(\mu_i^V - \nu_i\right)\right]$$

et

$$p\big(\Sigma_i^V \,|\alpha_i, \beta_i, J_i\big) \propto \frac{1}{\det(\Sigma_i^V)^{\beta_i}} \exp\left[-\alpha_i \mathrm{Tr}\Big(\left(\Sigma_i^V\right)^{-1} J_i\Big)\right]$$

où $\mathrm{Tr}(\cdot)$ est la trace d'une matrice. Ainsi, l'*a priori* utilisé sur les paramètres vidéo s'écrit : $p(\mu_i^V, \Sigma_i^V) = p(\mu_i^V|\Sigma_i^V)\, p(\Sigma_i^V)$. En annulant les dérivées partielles de (B.19) par rapport à μ_i^V et l'inverse de Σ_i^V on obtient respectivement

$$\sum_{t=1}^{T} \left(\Sigma_i^V\right)^{-1}\left(\mathbf{v}(t) - \mu_i^V\right)\, p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big) \underbrace{-\eta_i \left(\Sigma_i^V\right)^{-1}\left(\mu_i^V - \nu_i\right)}_{\frac{\partial}{\partial \mu_i^V} \ln p\left(\mu_i^V \left|\nu_i, \eta_i^{-1}\Sigma_i^V\right.\right)} = 0$$

et

$$\begin{aligned} &\frac{1}{2}\sum_{t=1}^{T}\left[\Sigma_i^V - \left(\mathbf{v}(t) - \mu_i^V\right)\left(\mathbf{v}(t) - \mu_i^V\right)^T\right]\, p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big) + \\ &\quad \underbrace{\frac{1}{2}\left[\Sigma_i^V - \eta_i\left(\mu_i^V - \nu_i\right)\left(\mu_i^V - \nu_i\right)^T\right]}_{\frac{\partial}{\partial\left(\Sigma_i^V\right)^{-1}} \ln p\left(\mu_i^V \left|\nu_i, \eta_i^{-1}\Sigma_i^V\right.\right)} + \underbrace{\beta_i \Sigma_i^V - \alpha_i\, J_i}_{\frac{\partial}{\partial\left(\Sigma_i^V\right)^{-1}} \ln p\left(\Sigma_i^V \left|\alpha_i, \beta_i, J_i\right.\right)} = 0. \end{aligned}$$

Finalement, les équations de mise à jour des paramètres vidéo sont données par

$$\left(\mu_i^V\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \mathbf{v}(t)\, p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big) + \eta_i\, \nu_i}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}) + \eta_i} \tag{B.22}$$

et

$$\begin{aligned} \left(\Sigma_i^V\right)^{(k+1)} = &\frac{\sum_{t=1}^{T}\left(\mathbf{v}(t) - \left(\mu_i^V\right)^{(k+1)}\right)\left(\mathbf{v}(t) - \left(\mu_i^V\right)^{(k+1)}\right)^T p\big(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\big)}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}) + 1 + 2\beta_i} \\ &+ \frac{\eta_i\left(\left(\mu_i^V\right)^{(k+1)} - \nu_i\right)\left(\left(\mu_i^V\right)^{(k+1)} - \nu_i\right)^T + 2\alpha_i\, J_i}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}) + 1 + 2\beta_i} \end{aligned} \tag{B.23}$$

Sans information *a priori* sur les vecteurs des valeurs moyennes μ_i^V, on opte alors un *a priori* de la forme $p(\Sigma_i^V) = \int p(\mu_i^V, \Sigma_i^V) d\mu_i$. Dans ces conditions, on obtient pour équations de mise à jour des paramètres vidéo les relations suivantes

$$\left(\mu_i^V\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \mathbf{v}(t)\, p\left(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\right)}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)})} \tag{B.24}$$

et

$$\left(\Sigma_i^V\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \left(\mathbf{v}(t) - \left(\mu_i^V\right)^{(k+1)}\right)\left(\mathbf{v}(t) - \left(\mu_i^V\right)^{(k+1)}\right)^T p\left(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\right) + 2\alpha_i\, J_i}{\sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}) + 2\beta_i}. \tag{B.25}$$

Il est intéressant de noter que l'*a priori* sur la matrice de covariance du $i^{\text{ème}}$ noyau est contenu dans la matrice J_i symétrique définie positive, α_i et β_i étant deux paramètres positifs de facteurs d'échelle. Le mode de la loi *a priori* $p(\Sigma_i^V)$ est obtenu en $(\alpha_i/\beta_i)\, J_i$ qui est donc la matrice de covariance la plus probable *a priori*.

B.2.3 Mise à jour des paramètres audio

L'annulation de la dérivée de l'équation (B.19) par rapport aux paramètres de localisation $\Gamma_i^A(f)$ fournit l'équation de mise à jour des paramètres audio. En adoptant un *a priori* conjugué pour $\Gamma_i^A(f)$, on obtient une distribution de Wishart inverse

$$\Gamma_i^A(f) \sim \mathcal{W}i(\alpha_i', \beta_i', J_i'(f)).$$

En annulant la dérivée de (B.19) par rapport à $\Gamma_i^A(f)$, on obtient

$$\frac{\partial \Psi(\theta)}{\partial \Gamma_i^A(f)} = \sum_{t=1}^{T} \left[-\frac{1}{\Gamma_i^A(f)} + \frac{\left(e^{a(t,f)}\right)^2}{2\left(\Gamma_i^A(f)\right)^2} \right] p\left(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\right) - \frac{\beta_i'}{\Gamma_i^A(f)} + \frac{\alpha_i'\, J_i'(f)}{\left(\Gamma_i^A(f)\right)^2}$$

donc

$$\left(\Gamma_i^A(f)\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \left(e^{a(t,f)}\right)^2 p\left(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\right) + 2\alpha_i'\, J_i(f)}{2\ \sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}) + 2\beta_i'}. \tag{B.26}$$

L'absence d'information *a priori* sur les profils spectraux des noyaux nous conduit à ne pas recourir à de loi *a priori* et donc les équations de mise à jour des paramètres de localisation des noyaux sont données par

$$\left(\Gamma_i^A(f)\right)^{(k+1)} = \frac{\sum_{t=1}^{T} \left(e^{a(t,f)}\right)^2 p\left(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)}\right)}{2\ \sum_{t=1}^{T} p(i|\mathbf{a}(t), \mathbf{v}(t), \Theta^{(k)})}. \tag{B.27}$$

Table des figures

Liste des tableaux

Bibliographie

[1] Frédéric Abrard et Yannick Deville : A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7):1389–1403, July 2005.

[2] Benoit Albouy et Yannick Deville : Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, April 2003.

[3] Benoit Albouy et Yannick Deville : Méthode temps-fréquence de séparation aveugle de sources basée sur la fonction de cohérence segmentée. *In Proc. GRETSI'03*, 2003.

[4] Shun-ichi Amari et Jean-François Cardoso : Blind Source Separation–Semiparametric Statistical Approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, November 1997.

[5] Shun-ichi Amari et Andrzej Cichocki : *Adaptive Blind Signal and Image Processing, Learning Algorithms and Applications.* Wiley, 2002.

[6] Jörn Anemüller et Birger Kollmeier : Amplitude modulation decorrelation for convolutive blind source separation. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 212–220, Helsinki, Finland, June 2000.

[7] B. Ans, J. Hérault et C. Jutten : Adaptive neural architectures : Detection of primitives. *In COGNITIVA*, pages 593–597, Paries, France, June 1985.

[8] Massoud Babaie-Zadeh, Christian Jutten et Ali Mansour : Sparse ICA via cluster-wise PCA. *Neurocomputing*, 69(13–15):1458–1466, August 2006.

[9] Massoud Babaie-Zadeh, Ali Mansour, Christian Jutten et Farrokh Marvasti : A Geometric Approach for Separating Several Speech Signals. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 798–806, Granada, Spain, 2004.

[10] Jon P. Barker et Frédéric Berthommier : Estimation of speech acoustics from visual speech features : a compararison of linear and non-linear models. *In Audio-Visual Speech Proc. (AVSP)*, pages 112–117, Santa Cruz, USA, 1999.

[11] William H.A. Beaudot : *Le traitement neuronal de l'information dans la rétine des vertébrés. Un creuset d'idées pour la vision artificielle.* Thèse de doctorat, Institut National Polytechnique de Grenoble, December 1994.

[12] Anthony J. Bell et Terrence J. Sejnowski : An information-maximization approach to blind source separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[13] A. Belouchrani, K. Abed-Meraim, M. G. Amin et A. M. Zoubir : Blind Separation of Nonstationary Sources. *IEEE Signal Processing Letters*, 11(7): 605–608, July 2004.

[14] Adel Belouchrani, Karim Abed-Meraim et Jean-François Cardoso : A blind source separation technique using second-order statistic. *IEEE Transactions on Signal Processing*, 45(2):434–444, February 1997.

[15] Adel Belouchrani et Moeness G. Amin : Blind source separation based on time frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, November 1998.

[16] Laurent Benaroya : *Séparation de plusieurs sources sonores avec un seul microphone*. Phd thesis, traitement du signal, Université de Rennes 1, June 2003.

[17] Laurent Benaroya, Frédéric Bimbot et Rémi Gribonval : Audio Source Separation With a Single Sensor. 14(1):191–199, January 2006.

[18] Alexandre Benoit et Alice Caplier : Head nods analysis : interpretation of non verbal communication gestures. *In ICIP*, Genova, Italy, 2005.

[19] Alexandre Benoit et Alice Caplier : Motion Estimator Inspired From Biological Model For Head Motion Interpretation. *In WIAMIS*, Montreux, Switzrland, 2005.

[20] Christian Benoît, Tahar Mohamadi et Sonia Kandel : Effects of phonetic context on audio-visual intelligibility of French. *J. Speech and Hearing Research*, 37:1195–1293, 1994.

[21] Christian Benoît, Thierry Guiard-Marigny, Bertrand Le Goff et Ali Adjoudani : Which components of the face humans and machines best speechread ? *In* D.G. Stork & M.E. Hennecke, éditeur : *Speechreading by man and machine : Models, Systems and Applications*, pages 315–328. Springer-Verlag, New York, 1996.

[22] Lynne E. Bernstein, Edward T. Jr. Auer et Sumiko Takayanagi : Auditory speech detection in noise enhanced by lipreading. *Speech Comm.*, 44(1–4):5–18, 2004.

[23] Lynne E. Bernstein et Christian Benoît : For speech perception by humans or machines, three senses are better than one. *In Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pages 1477–1480, Philadelphia, USA, 1996.

[24] David Burshtein et Sharon Gannot : Speech Enhancement Using a Mixture-Maximum Model. *IEEE Transactions on Speech and Audio Processing*, 10(6):341–351, September 2002.

[25] Véronique Capdevielle, Christine Servière et Jean-Louis Lacoume : Blind Separation of Wide-Band Sources : Application to rotating machine signals. *In 8th European Signal Processing Conference*, volume 3, pages 2085–2088, Trieste, Italy, 1996.

[26] Jean-François CARDOSO : Infomax and Maximum Likelihood for Blind Source Separation. *IEEE Signal Processing Letters*, 4(4):112–114, April 1997.

[27] Jean-François CARDOSO : High-Order Contrasts for Independent Component Analysis. *Neural Computation*, 11:157–192, 1999.

[28] Jean-François CARDOSO : Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, October 1998.

[29] Jean-François CARDOSO et Beate Hvam LAHELD : Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

[30] Jean-François CARDOSO et Dinh-Tuan PHAM : Séparation par l'indépendance et la parcimonie. *In GRETSI*, Paris, France, September 2003.

[31] Jean-François CARDOSO et Antoine SOULOUMIAC : Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[32] Nabil CHARKANI et Yannick DEVILLE : A convolutive source separation method with self-optimizing non-linearities. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2909–2912, Phoenix, USA, April 1999.

[33] Eric CHAUMETTE, Pierre COMMON et Daniel MULLER : ICA-based technique for radiating sources estimation : application to airport surveillance. *In International Workshop on High Order Statistics*, pages 210–214, South Lake Tahoe, USA, June 1993.

[34] Yong Duk CHO et Ahmet KONDOZ : Analysis and Improvement of s Statistical Model-Based Voice Activity Detector. *IEEE Signal Processing Letters*, 8(10): 276–278, October 2001.

[35] Seungjin CHOI et Andrzej CICHOCKI : Blind Separation of nonstationary sources in noisy mixtures. *Electronics letters*, 36(9):848–849, April 2000.

[36] Andrzej CICHOCKI, Rolf UNBEHAUEN et E. RUMMERT : Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387, August 1994.

[37] Israel COHEN : Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator. *IEEE Transactions on Signal Processing*, 9(4):113–116, April 2002.

[38] Pierre COMON : Independent component analysis, a new concept ? *Signal Processing*, 36(3):287–314, April 1994.

[39] T. M. COVER et J. A. THOMAS : *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[40] Adriana DAPENA, Mónica F. BUGALLO et Luis CASTEDO : Separation of convolutive mixtures of temporally-white signals : a novel frequency-domain approach. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 315–320, San Diego, USA, December 2001.

[41] G. DARMOIS : Analyses des liaisons de probabilité. *In Int. Statistics Conferences 1947*, volume III A, page 231, Washington (D.C.), 1951.

[42] G. Darmois : Analyse Générale des Liaisons Stochastiques. *Rev. Inst. Internat. Stat.*, 21:2–8, 1953.

[43] L. De Lathauwer, D. Callaerts, B. De Moor et J. Vandewalle : Fetal electrocardiogram extraction by source subspace separation. *In Proc. IEEE Workshop on HOS*, pages 134–138, Girona, Spain, June 12–14 1995.

[44] Serge Dégerine et Ahmed Zaidi : Maximum de vraisemblance exact pour la séparation aveugle d'un mélange instantané de sources autorégressives gaussiennes. *In GRETSI*, volume 4, pages 1141–1144, Vannes, France, September 1999.

[45] Sabine Deligne, Gerasimos Potamianos et Chapalathy Neti : Audio-Visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization). *In Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pages 1449–1452, 2002.

[46] A. P. Dempster, N. M. Laird et D. B. Rubin : Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B.*, 39:1–38, 1977.

[47] Frédéric Desobry et Cédric Févotte : Kernel PCA based estimation of the mixing matrix in linear instantaneous mixtures of sparse sources. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 669–672, Toulouse, France, May 2006.

[48] David L. Donoho : On minimum entropy deconvolution. *In Proc. 2nd Applied Time Series Symp.*, Tulsa, 1980. reprinted in Applied Time Series Analysis II, Academic Press, New York, 1981, pp. 565-609.

[49] Stéphane Dupont et Juergen Luettin : Audio-Visual Speech Modeling for Continuous Speech Recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, September 2000.

[50] Yariv Ephraim et Israel Cohen : Recent Advancements in Speech Enhancement. Rapport technique, George Mason University, 2004.

[51] Yariv Ephraim et David Malah : Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, December 1984.

[52] Norman P. Erber : Interaction of audition et vision in the recognition of oral speech stimuli. *J. Speech and Hearing Research*, 12:423–425, 1969.

[53] D.K. Freeman, G. Cosier, C.B. Southcott et I. Boyd : The voice activity detector for pan-european digital cellular mobile telephone service. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 369–372, 1989.

[54] M. Gaeta et J.-L. Lacoume : Source separation without a priori knowledge : the maximum likelihood solution. *In EUSIPCO*, volume 2, pages 621–641, Barcelone, Spain, September 1990.

[55] Saeed Gazor et Wei Zhang : A Soft Voice Activity Detector Based on a Laplacian-Gaussian Model. *IEEE Transactions on Speech and Audio Processing*, 11(5):498–505, September 2003.

[56] Saeed GAZOR et Wei ZHANG : Speech Probability Distribution. *IEEE Signal Processing Letters*, 10(7):204–207, July 2003.

[57] Allen. GERSHO et Robert M. GRAY : *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Amsterdam, 1992.

[58] Laurent GIRIN : Joint Matrix Quantization of Face Parameters and LPC Coefficients for Low Bit Rate Audiovisual Speech Coding. *IEEE Transactions on Speech and Audio Processing*, 12(3):265–276, May 2004.

[59] Laurent GIRIN, Jean-Luc SCHWARTZ et Gang FENG : Audio-visual enhancement of speech in noise. *J. Acoust. Soc. Am.*, 109(6):3007–3020, June 2001.

[60] Roland GOECKE, Gerasimos POTAMIANOS et Chalapathy NETI : Noisy audio feature enhancement using audio-visual speech data. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2025–2028, Orlando, USA, May 2002.

[61] Ken W. GRANT et Philip-Franz SEITZ : The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, 108: 1197–1208, 2000.

[62] Rémi GRIBONVAL : Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, USA, 2002.

[63] Rémi GRIBONVAL et Sylvain LESAGE : A survey of Sparse Component Analysis for Blind Source Separation : principles, perspectives, and new challenges. *In ESANN (European Symposium on Artificial Neural Networks), Advances in Computational Intelligence and Learning*, pages 323–330, Bruges, April 2006.

[64] Nathalie GUYADER : *Perception de scènes et caractérisation d'objets en fonction du contexte : taille, orientation, perspective.* Thèse de doctorat, Institut National Polytechnique de Grenoble, July 2004.

[65] W. HÄRDLE : *Smoothing Techniques, with implementation in S.* Springer-Verlag, 1990.

[66] J. HÉRAULT, C. JUTTEN et B. ANS : Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétrique en apprentissage non supervisé. *In Proc. GRETSI'85*, volume 2, pages 1017–1020, Nice, France, May 1985.

[67] Jeanny HÉRAULT et William BEAUDOT : Motion Processing in the Retina : about a Velocity Matched Filter. pages 129–136, Brussels, Belgium, April 1993.

[68] Jeanny HÉRAULT et Christian JUTTEN : Space or time adaptive signal processing by neural networks models. *In Intern. Conf. on Neural Networks for Computing*, pages 206–211, Snowbird, USA, 1986.

[69] Aapo HYVÄRINEN, Juha KARHUNEN et Erkki OJA : *Independent Component Analysis.* Wiley, New York, 2001.

[70] Aapo HYVÄRINEN et Erkki OJA : A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[71] Aapo Hyvarinen et Erkki Oja : Independent Component Analysis : Algorithms and Applications. *Neural Networks*, 13(4–5):411–430, June 2000.

[72] A. Jourjine, S. Rickard et O. Yilmaz : Blind separation of disjoint orthogonal signals : demixing N sources from 2 mixtures. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 2985–2988, Istanbul, Turkey, June 2000.

[73] C. Jutten, L. Nguyen Thi, E. Dijkstra, E. Vittoz et J. Caelen : Blind separation of sources : an algorithm for separation of convolutive mixtures. *In International Signal Processing Workshop on Higher Order Statistics*, pages 273–276, Chamrousse, France, July 1991.

[74] Christian Jutten : Algorithmes fondés sur l'information mutuelle. *In Proc. école de printemps de Villard-de-Lans : de la séparation de sources à l'analyse en composantes indépendantes*, pages 145–176, Villard-de-Lans, May 2001.

[75] Christian Jutten et Rémi Gribonval : L'analyse en composantes indépendantes : un outil puissant pour le traitement de l'information. *In Proc. GRETSI'03*, Paris, France, 2003.

[76] Christian Jutten et Jeanny Hérault : Blind separation of sources. Part I : An adaptive algorithm based on a neuromimetic architecture. *Signal Processing*, 24(1):1–10, July 1991.

[77] Christian Jutten et Anisse Taleb : Source separation : from dusk till dawn. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 15–26, Helsinki, Finland, June 2000.

[78] Allan Kardec Barros, Ali Mansour et Noboru Ohnishi : Removing artifacts from electrocardiographic signals using independent components analysis. *Neurocomputing*, 22:173–186, 1998.

[79] Jeesun Kim et Davis Chris : Investigating the audio–visual speech detection advantage. *Speech Comm.*, 44(1–4):19–30, 2004.

[80] Tahar Lallouache : Un poste visage-parole. Acquisition et traitement des contours labiaux. *In Proc. Journées d'Etude sur la Parole (JEP) (French)*, Montréal, 1990.

[81] J. F. Lawless : Inference in the generalized gamma and log-gamma distribution. *Technometrics*, 22:67–82, 1980.

[82] Régine Le Bouquin-Jeannès et Gérard Faucon : Study of a voice activity detector and its influence on a noise reduction system. *Speech Comm.*, 16:245–254, 1995.

[83] Eric Le Carpentier et Cédric Févotte : Séparation de sources autorégressives gausiennes par maximum de vraisemblance et filtrage de Kalman. *In GRETSI*, volume 2, pages 323–326, Toulouse, France, September 2001.

[84] Ulf A. Lindgren et Holger Broman : Source separation using a criterion based on second-order statistics. *IEEE Transactions on Signal Processing*, 46(7):1837–1850, July 1998.

[85] Zied Malouche et Odile Macchi : Adaptive Unsupervised Extraction of One Component of a Linear Mixture with a Single Neuron. *IEEE Transactions on Neural Networks*, 9(1):123–138, January 1998.

[86] Rainer MARTIN et Colin BREITHAUPT : Speech Enhancement in the DFT Domain using Laplacian Speech Priors. *In Proc. IWAENC'03*, pages 87–90, Kyoto, Japan, September 2003.

[87] H. MCGURK et J. MCDONALD : Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[88] C.A. MEAD et M.A. MAHOWALD : A silicon model for early visual processing. *Neural Network*, 1(1):91–97, 1988.

[89] Fredy D. NEESER et James L. MASSEY : Proper complex random processes with applications to information theory. *IEEE Transactions on Information Theory*, 39(4):1293–1302, July 1993.

[90] Hoang-Lan NGUYEN-THI et Christian JUTTEN : Blind source separation for convolutive mixtures. *Signal Processing*, 45:209–229, 1995.

[91] Dirk ORMONEIT et Volker TRESP : Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9:639–650, 1998.

[92] Athanasios PAPOULIS : *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, third édition, 1991.

[93] Lucas PARRA et Clay SPENCE : Convolutive blind separation of non stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, May 2000.

[94] Eric D. PETAJAN : *Automatic lipreading to enhance speech recognition*. Phd. thesis, University of Illinois, 1984.

[95] Dinh-Tuan PHAM : Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. *Signal Processing*, 81(4):855–870, 2001.

[96] Dinh-Tuan PHAM : Joint approximate diagonalization of positive definite matrices. *SIAM J. Matrix Anal. And Appl.*, 22(4):1136–1152, 2001.

[97] Dinh-Tuan PHAM : Fast Algorithms for Mutual Information Based Independent Component Analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700, October 2004.

[98] Dinh-Tuan PHAM et Jean-François CARDOSO : Source adaptive blind source separation : Gaussian models and sparsity. *In Proceedings of SPIE*, volume 5207. Wavelets : Applications in Signal and Image Processing X, November 2003.

[99] Dinh-Tuan PHAM et Jean-François CARDOSO : Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, 49(9):1837–1848, September 2001.

[100] Dinh Tuan PHAM et Philippe GARAT : Blind Separation of Mixture of Independent Sources Through a Quasi-Maximum Likelihood Approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, July 1997.

[101] Dinh-Tuan PHAM, Philipe GARRAT et Christian JUTTEN : Separation of a mixture of indenpendent sources through a maximum likelihood approach. *In EUSIPCO*, pages 771–774, 1992.

[102] Dinh-Tuan PHAM, Christine SERVIÈRE et Hakim BOUMARAF : Blind separation of convolutive audio mixtures using nonstationarity. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, April 2003.

[103] Bernard PICINBONO : Second-order complex random vectors and normal distributions. *IEEE Transactions on Signal Processing*, 44(10):2637–2640, October 1996.

[104] Gerasimos POTAMIANOS, Chalapathy NETI, Guillaume GRAVIER, Ashutosh GARG et Andrew W. SENIOR : Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[105] Matthieu PUIGT et Yannick DEVILLE : Time-frequency ratio-based blind separation methods for attenuated and time-delayed sources. *Signal Processing*, 19:1348–1379, 2005.

[106] C. G. PUNTONET, A. PRIETO, C. JUTTEN, M. RODRÍGUEZ-ALVAREZ et J. ORTEGA : Separation of sources : A geometry-based procedure for reconstruction of n-valued signals. *Signal Porcessing*, 46:267–284, 1995.

[107] Douglas A. REYNOLDS et Richard C. ROSE : Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[108] Scott RICKARD, Radu BALAN et Justinian ROSCA : Real-time time-frequency based blind source separation. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, San Diego (USA), December 2001.

[109] Bertrand RIVET, Laurent GIRIN et Christian JUTTEN : Solving the indeterminations of blind source separation of convolutive speech mixtures. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 533–536, Philadelphia, USA, March 2005.

[110] Bertrand RIVET, Laurent GIRIN et Christian JUTTEN : Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Submitted to Speech Com.*, 2006.

[111] Bertrand RIVET, Laurent GIRIN et Christian JUTTEN : Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, (Accepted for publication).

[112] Bertrand RIVET, Laurent GIRIN et Christian JUTTEN : Log-Rayleigh distribution : a simple and efficient statistical representation of log-spectral coefficients. *IEEE Transactions on Speech and Audio Processing*, (Accepted for publication) 2006.

[113] Jordi ROBERT-RIBES, Jean-Luc SCHWARTZ, Tahar LALLOUACHE et Pierre ESCUDIER : Complementarity and synergy in bimodal speech : Auditory, visual, and audio-visual identification of French oral vowels in noise. *J. Acoust. Soc. Am.*, 103(6):3677–3689, 1998.

[114] Peter SCHREIER et Louis SCHARF : Second-order analysis of improper complex random vectors and processes. *IEEE Transactions on Signal Processing*, 51(3): 714–725, March 2003.

[115] Christine SERVIÈRE et Dinh-Tuan PHAM : A novel method for permutation correction in frequency-domain in blind separation of speech mixtures. *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 807–815, Granada, Spain, 2004.

[116] Hichem SNOUSSI et Ali MOHAMMAD-DJAFARI : Penalized maximum likelihood for multivariate Gaussian mixture. *In Proc. Bayesian Inference and Maximum Entropy Methods. R. L. Fry, Ed. MaxEnt Workshops*, pages 36–46, August 2001.

[117] David SODOYER : *La séparation de sources audiovisuelles.* Thèse de doctorat, Institut National Polytechnique Grenoble, 2004.

[118] David SODOYER, Laurent GIRIN, Christian JUTTEN et Jean-Luc SCHWARTZ : Developing an audio-visual speech source separation algorithm. *Speech Comm.*, 44(1–4):113–125, October 2004.

[119] David SODOYER, Bertrand RIVET, Laurent GIRIN, Jean-Luc SCHWARTZ et Christian JUTTEN : An analysis of visual speech information applied to voice activity detection. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 601–604, Toulouse, France, 2006.

[120] David SODOYER, Jean-Luc SCHWARTZ, Laurent GIRIN, Jacob KLINKISCH et Christian JUTTEN : Separation of audio-visual speech sources : a new approach exploiting the audiovisual coherence of speech stimuli. *Eurasip Journal on Applied Signal Processing*, 2002(11):1165–1173, 2002.

[121] Jongseo SOHN, Nam Soo KIM et Wonyong SUNG : A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999.

[122] Murray R. SPIEGEL : *Schaum's Mathematical Handbook of Formulas & Tables.* McGraw-Hill.

[123] David G. STORK et Mickael E. HENNECKE : *Speechreadings by Humans and Machines.* Berlin, Germany : Springer-Verlag, 1996.

[124] Yannis STYLIANOU, Olivier CAPPÉ et Eric MOULINES : Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, March 1998.

[125] W.H. SUMBY et I. POLLACK : Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, 26:212–215, 1954.

[126] Quentin SUMMERFIELD : Some preliminaries to a comprehensive account of audio-visual speech perception. *In* B. DODD et R. CAMPBELL, éditeurs : *Hearing by Eye : The Psychology of Lipreading*, pages 3–51. Lawrence Erlbaum Associates, 1987.

[127] Anisse TALEB et Christian JUTTEN : Source separation in post nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, October 1999.

[128] S. Gökhun Tanyer et Hamza Özer : Voice Activity Detection in Nonstationary Noise. *IEEE Transactions on Speech and Audio Processing*, 8(4):478–482, July 2000.

[129] Pascal Teissier, Jordi Rober-Ribes, Jean-Luc Schwartz et Anne Guérin-Dugué : Comparing Models for Audiovisual Fusion in a Noisy-Wovel Recognition Task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, November 1999.

[130] Kari Torkkola : Blind separation for audio signals : are we there yet ? *In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 239–244, Aussois, France, 1999.

[131] Jitendra E. Tugnait : Blind estimation of digital communication channel impluse response. *IEEE Transactions on Communications*, 42(2-4):1606–1616, 1994.

[132] Nathalie Vallée : *Systèmes vocaliques : de la typologie aux prédictions*. Thèse de Doctorat en sciences du langage, Université Stendhal, Grenoble, 1994.

[133] Nathalie Vallée, Louis-Jean Boë et Yohan Payan : Vowel Prototypes for UPSID'S 33 Phonemes. *In International Congress of Phonetic Sciences*, volume 1–4, pages 424–427, Stockholm, 1995.

[134] Wenwu Wang, Darren Cosker, Yulia Hicks, Saied Sanei et Jonathon A. Chambers : Video assisted speech source separation. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[135] Wenwu Wang, Saeid Sanei et Jonathon A. Chambers : Penalty Function-Based Joint Diagonalization Approach for Convolutive Blind Separation of Nonstationary Sources. *IEEE Transactions on Signal Processing*, 53(5):1654–1669, May 2005.

[136] Ehud Weinstein, Mier Feder et Alan V. Appenheim : Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, 1(4):405–413, October 1993.

[137] Hani Yehia, Philip Rubin et Eric Vatikiotis-Bateson : Quantitative association of vocal-tract and facial behavior. *Speech Comm.*, 26(1):23–43, 1998.

[138] Daniel Yellin et Ehud Weinstein : Criteria for multichannel signal separation. *IEEE Transactions on Signal Processing*, 42(8):2158–2168, August 1994.

[139] Özgür Yilmaz et Scott Rickard : Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, 52(7): 1830–1847, July 2004.

[140] Vicente Zarzoso et Asoke K. Nandi : Noninvasive fetal electrocardiogram extraction : blind source separation versus adaptative noise cancellation. *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, January 2001.

[141] E. Zwicker et U. Zwicker : *Psychoacoustics, Facts and Models*. Springer-Verlag, Berlin, Germany, 1990.

Résumé

Cette thèse est dédiée à la modélisation conjointe des modalités audio et vidéo de la parole et à son exploitation pour la séparation de sources. Tout d'abord, une modélisation probabiliste bimodale de la parole audiovisuelle à base de mélange de noyaux est proposée. Cette modélisation est ensuite exploitée pour la détection des silences. De plus, nous proposons une détection purement visuelle des silences en s'appuyant sur l'observation des lèvres du locuteur. Ce dernier procédé présente l'avantage d'être indépendant d'un bruit acoustique. Ces deux modélisations sont ensuite exploitées pour la séparation de mélanges convolutifs de sources audiovisuelles. Nous résolvons ainsi le problème classique des indéterminations des méthodes de séparation dans le domaine fréquentiel avant de proposer une méthode géométrique qui utilise les périodes de silence de la source d'intérêt. Les algorithmes proposés sont validés par des expériences sur des corpus multi-locuteurs et multi-langues.

Abstract

This thesis is dedicated to both the joint modeling of the audio and visual modalities of speech and its use in source separation. A mixture of kernels is first proposed to model the bi-modality of audiovisual speech. This modeling is then exploited to detect the silence phases of speech. Moreover, we propose a purely visual detection of silence based on the lip movements of the speaker. The later detection is robust to any acoustic environment. These two modelings are then exploited in source separation of convolutive mixtures. We first solve the classical indeterminacies encountered by frequency domain separation algorithms. We then propose a geometric separation which exploits the silence of the source of interest. The proposed algorithms are validated by experiments on multi-speakers and multi-languages databases.