MASTER 2 SCIENCES DE L'INGÉNIEUR
Université Pierre et Marie Curie - Paris 6

PARCOURS ATIAM

RAPPORT DE STAGE DE FIN D'ÉTUDE

Adaptation au locuteur pour la séparation de la parole par NMF

Auteur
Guillaume DORAS

Lieu du stage
Ircam - Analyse et synthèse des sons

Responsable
Nicolas OBIN - Ircam
(Prénom.Nom@ircam.fr)

28 juillet 2016

Résumé

Ce rapport de stage de 2ème année de master a pour objet l'utilisation de la NMF semi-supervisée pour la séparation de sources audio, et en particulier de la parole. La contribution principale de ce stage est une méthode de séparation de la parole par adaptation à un locuteur inconnu de l'apprentissage réalisé sur un autre locuteur connu a priori.
Tout d'abord, un rappel sur les méthodes de séparation de parole par NMF semi-supervisée - et leurs limites actuelles dans le cas d'un locuteur inconnu - est proposé, ainsi qu'un état de l'art des améliorations envisagées jusqu'à présent dans la littérature. Après avoir discuté de ces différentes approches, une nouvelle méthode d'adaptation à un locuteur inconnu de l'apprentissage réalisé sur un autre locuteur sera proposée, ainsi que plusieurs contraintes permettant d'améliorer la qualité de la séparation. Enfin, le modèle d'adaptation proposé et les différentes contraintes sont évaluées et comparées aux résultats obtenus sans adaptation au locuteur.

Mots-clés : Séparation de la parole, débruitage de la parole, factorisation en matrices non-négatives (NMF), modèle source/filtre, NMF sous contraintes, adaptation au locuteur, déformation spectrale. //

Abstract

This master thesis is on the use of semi-supervised NMF for audio sources separation, and in particular speech separation. The main contribution of this work is a speech separation method based on an adaptation to an unknown speaker of a prior training performed with a known speaker.
First, an overview of the speech separation by semi-supervised NMF methods - and their current limits in the case of an unknown speaker - is presented, as well as a state-of-the-art of improvements proposed until now. After discussing those different approaches, a new adaptation method to an unknown speaker of the training performed with a known speaker is presented, as well as several constraints aimed at improving the separation quality. Finally, the proposed adaptation model and the different constraints are evaluated and compared to the results obtained without speaker adaptation.

Keywords : Speech separation, speech enhancement, non-negative matrix factorization (NMF), source/filter model, constrained NMF, speaker adaptation, spectral deformation. //

Table des matières

Introduction **1**

1 Etat de l'art en séparation de la parole par NMF **2**

1.1 Rappels sur la séparation de sources par NMF 2

1.1.1 Généralités sur la NMF et son utilisation en analyse audio 2

1.1.2 Application à la séparation de sources semi-supervisée 3

1.2 Séparation semi-supervisée de la parole 4

1.2.1 Rappels sur le modèle source-filtre 4

1.2.2 Formulation du modèle source/filtre dans un cadre NMF 5

1.2.3 Positionnement du problème 6

1.3 Déformation des bases spectrales 8

1.3.1 Le modèle de déformation additive 8

1.3.2 Le modèle de déformation multiplicative 8

1.3.3 Interprétation des différentes déformations 9

2 Contributions **11**

2.1 Intérêt et limites des modèles existants pour le problème posé 11

2.1.1 Le modèle de déformation additif 11

2.1.2 Le modèle de déformation multiplicatif par matrice 12

2.2 Proposition d'un nouveau modèle de déformation 13

2.2.1 Le modèle de déformation multiplicative par tenseur 13

2.2.2 NMF des modèles de déformation tensorielle vs. multiplicative terme à terme 15

2.3 Application du modèle de déformation à la séparation de sources 17

2.3.1 Les contraintes de "petite déformation" 17

2.3.2 Les contraintes de "déformation cohérente" 18

3 Résultats **22**

3.1 Description du protocole de test 22

3.1.1 La base de test 22

3.1.2 Les métriques utilisées 22

3.1.3 Le protocole de test utilisé 23

3.2 Scores obtenus 24
3.2.1 Rappel des résultats pour l'algorithme sans déformation 24
3.2.2 Résultats pour l'algorithme avec déformation 25
3.2.3 Resultats avec la déformation et de multiples références 26
3.3 Discussion des résultats 26
3.3.1 Contraintes de petite déformation 26
3.3.2 Contrainte de réciprocité 27
3.3.3 Contrainte d'identité 27

Conclusion et perspectives **28**

Annexe A Calculs des gradients des modèles de déformation **29**
A.1 $\mathbf{V} = \mathbf{WH}$ 29
A.2 $\mathbf{V} = (\mathbf{W} + \mathbf{G})\mathbf{H}$ 30
A.3 $\mathbf{V} = (\mathbf{D} \otimes \mathbf{W})\mathbf{H}$ 31

Annexe B Calculs des gradients des contraintes utilisées **32**
B.1 Contraintes de petite déformation 32
B.1.1 Contrainte de petit gain 32
B.1.2 Contrainte de maximum de corrélation entre mêmes phonèmes 32
B.1.3 Contrainte de lissage spectral 33
B.2 Contraintes de déformation cohérente 33
B.2.1 Contraintes de décorrélation avec le bruit 33
B.2.2 Contrainte de réciprocité 34
B.2.3 Contraintes d'identité 35

Annexe C Equivalence de la NMF des déformations additive et multiplicative terme à terme **36**

Annexe D Propriété de la déformation par tenseur **38**
D.1 Cas $\mathscr{C}(\mathbf{D}) = \lambda \|\mathbf{D}\|_F^{\alpha}$ 38
D.2 Cas $\mathscr{C}(\mathbf{D}) = \lambda \|\mathbf{D}\|_1^{\alpha}$ 39

Annexe E Correspondance de la NMF des déformation tensorielle et multiplicative terme à terme **40**
E.1 Illustration de l'équivalence 42

Remerciements

Je tiens à remercier toute l'équipe pédagogique du master ATIAM de m'avoir donné l'opportunité de vivre cette année.

Je tiens également à remercier Nicolas Obin de m'avoir accepté pour ce stage et de m'avoir guidé et soutenu au cours des derniers mois.

Je tiens enfin à remercier Laurent Benaroya d'avoir eu la patience et la gentillesse de partager avec moi ses vastes connaissances sur la NMF, et Damien Bouvier pour son aide précieuse sur les algorithmes utilisés.

Notations

Mathématiques

$\mathbf{X}$	Matrice
$\mathbf{X}_{i,j}$	Coefficient d'indices i, j de la matrice $\mathbf{X}$
$\mathbf{x}_j$	$j^{\text{ème}}$ colonne de la matrice $\mathbf{X}$
X_i	$i^{\text{ème}}$ ligne de la matrice $\mathbf{X}$
$\otimes$	Produit de Hadamard (terme à terme)
$(.)^T$	Transposé

Algorithmiques

$(.)^{(i)}$	Valeur à la $i^{\text{ème}}$ itération
t, T	Indice de trame temporelle, nombre de trames temporelles
f, F	Indice de point fréquentiel, nombre de points fréquentiels
k, b, K	Indice de composante élémentaire, nombre de composantes élémentaires
$\mathbf{X}$	Représentation temps-fréquence à valeurs positives ou nulles observée
$\mathbf{V}$	Représentation temps-fréquence à valeurs positives ou nulles estimée
$\mathbf{W}$	Matrice de dictionnaires spectraux à valeurs positives ou nulles
$\mathbf{H}$	Matrice d'activations temporelles à valeurs positives ou nulles
$\mathbf{D}, \mathbf{\Delta}_b$	Matrice de déformation à valeurs positives ou nulles
$\mathbf{\Delta}$	Tenseur de déformation à valeurs positives ou nulles

Acronymes

NMF	Factorisation en matrices non-négatives
EUC	Euclidienne, distance euclidienne
KL	Kullback-Leibler, divergence de Kullback-Leibler
IS	Itakura-Saito, divergence d'Itakura-Saito
TFCT	Transformée de Fourier à court terme

Introduction

Ce stage a été proposé dans le cadre d'un projet de recherche plus vaste mené au sein de l'équipe Analyse et Synthèse des Sons de l'Ircam et que l'on désigne par l'analyse de scènes sonores, c'est-à-dire la détection, la localisation et la séparation de sources sonores. C'est à cette problématique de séparation de sources sonores que nous nous intéresserons ici, et en particulier au cas où la source à séparer est une voix parlée dans un environnement sonore inconnu a priori.

Le sujet du stage s'inscrit dans la continuité de celui mené l'an dernier au sein de l'équipe, et qui avait permis la mise au point d'un algorithme de séparation de parole par NMF semi-supervisée à partir d'une modélisation source/filtre de la voix et d'un apprentissage préalable phonèmes par phonème des bases spectrales du filtre pour un locuteur donné [Bouvier, 2015].

Cette approche donne de bons résultats - similaires à ceux de l'état de l'art pour les méthodes de séparation supervisées - dans le cas où la séparation se fait sur le même locuteur que celui qui a servi à l'apprentissage. Les résultats sont moins bons dès lors que le locuteur cible n'est pas celui pour lequel l'apprentissage a eu lieu. Or, il n'est pas toujours possible en pratique de pouvoir réaliser un apprentissage a priori sur un locuteur cible : il est donc apparu nécessaire de chercher à rendre plus robuste cet algorithme dans les cas où la parole à séparer est celle d'un locuteur pour lequel aucune information n'est disponible a priori.

L'idée mise en oeuvre ici repose sur l'hypothèse que les enveloppes spectrales d'un même phonème ne présentent que de faibles déformations d'un locuteur à l'autre. L'objectif du stage était donc de proposer un modèle de déformation des bases spectrales entre un locuteur de référence et un locuteur cible s'exprimant dans un environnement sonore inconnu a priori d'une part, et de proposer une méthode d'estimation de cette déformation à la volée au cours de la séparation par NMF d'autre part.

Après avoir brièvement rappelé le formalisme de la NMF et son adaptation au modèle source/filtre, nous soulignerons dans une première partie les limites de cette approche pour la séparation d'un signal de parole d'un locuteur inconnu, et formulerons le problème de l'adaptation au locuteur auquel nous nous intéresserons dans la suite de ce rapport.
Après avoir dressé un état de l'art des approches actuellement disponibles pour réaliser une adaptation entre locuteurs, nous analyserons dans une deuxième partie leurs limites pour le problème posé. Nous proposerons alors un nouveau modèle de déformation plus adéquat et différentes contraintes devant permettre d'améliorer la qualité de la séparation.
Nous présenterons enfin dans une troisième partie les résultats obtenus avec ce nouveau modèle et les différentes contraintes utilisées, que nous comparerons à ceux obtenus sans adaptation au locuteur. Nous pourrons alors en conclure que la méthode proposée répond au problème posé de manière satisfaisante.

1 Etat de l'art en séparation de la parole par NMF

Nous allons dans un premier temps rappeler les principes de la NMF et de son utilisation en séparation de sources, ainsi que le formalisme du modèle source/filtre utilisé dans ce cadre. Nous rappelerons ensuite les résultats actuels obtenus en séparation de parole avec cette méthode - résultats qui nous amèneront à formuler explicitement le problème que nous traiterons dans la suite du rapport. Nous évoquerons alors différentes approches de ce problème déjà envisagées dans la littérature.

1.1 Rappels sur la séparation de sources par NMF

1.1.1 Généralités sur la NMF et son utilisation en analyse audio

On représente classiquement un signal sonore par son spectrogramme, c'est-à-dire la répartion de son énergie en fonction du temps et de la fréquence. Il s'exprime mathématiquement sous la forme d'une matrice que l'on notera **X** par la suite, et qui est obtenue à partir de la TFCT de ce signal. On notera ses dimensions $F \times T$, où F est le nombre de bins fréquentiels et T le nombre de bins temporels de la TFCT.

Cette représentation temps-fréquence du signal sonore permet d'envisager sa factorisation en matrices non-négatives. On désigne cette factorisation par son acronyme anglo-saxon NMF (non-negative matrix factorization), et par extension l'algorithme itératif qui permet d'en fournir une estimation **V** à partir du spectrogramme **X** du signal observé [Févotte et al., 2009], ce qui s'écrit classiquement :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{WH} \tag{1}$$

La matrice **W** est une matrice de dimensions $F \times K$ qui représentent les K bases spectrales du signal, et la matrice **H** est une matrice de dimensions $K \times T$ qui représente les activations temporelles des K différentes bases.

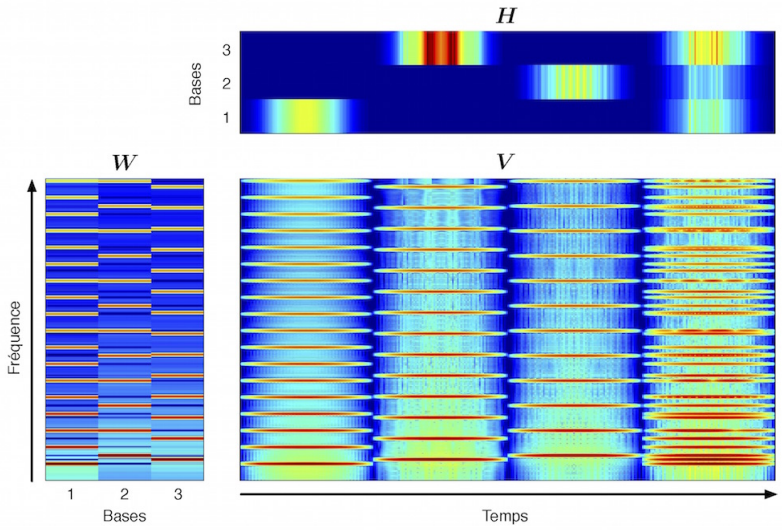


FIGURE 1 – Réprésentation d'une décomposition NMF d'un spectrogramme d'un signal de piano, avec $K = 3$ bases spectrales (Do, mi, sol)

Le produit **V** de ces deux matrices est une estimation du spectogramme observé **X** que l'on obtient en

minimisant une β-divergence entre $\mathbf{X}$ et $\mathbf{V}$ notée ici $d_\beta(\mathbf{X}|\mathbf{V})$, et qui s'exprime sous la forme générale :

$$d_\beta(\mathbf{X}|\mathbf{V}) = \sum_{i,j} \frac{1}{\beta(\beta-1)} (\mathbf{X}_{ij}^\beta + (\beta-1)\mathbf{V}_{ij}^\beta - \beta \mathbf{X}_{ij}\mathbf{V}_{ij}^{\beta-1})$$

Les valeurs de $\mathbf{W}$ et $\mathbf{H}$ qui minimisent cette β-divergence sont obtenues par une descente de gradient dont le détail des calculs est rappelé en annexe A. On retiendra en particulier que, grâce à la non-négativité des matrices $\mathbf{W}$ et $\mathbf{H}$, le pas de cette descente de gradient peut être exprimé de façon à obtenir des règles de mise à jour multiplicative pour $\mathbf{W}$ et $\mathbf{H}$ [Lee and Seung, 1999] :

$$\begin{cases} \mathbf{W}^{(n+1)} \longleftarrow \mathbf{W}^{(n)} \otimes \dfrac{(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T}{\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T} \\ \mathbf{H}^{(n+1)} \longleftarrow \mathbf{H}^{(n)} \otimes \dfrac{\mathbf{W}^T(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})}{\mathbf{W}^T\mathbf{V}^{\cdot(\beta-1)}} \end{cases} \tag{2}$$

On utilisera par la suite cette méthode de mise à jour étendue aux cas où la fonction à minimiser est la somme de $d_\beta(\mathbf{X}|\mathbf{V})$ et d'un coût additionnel $\mathscr{C}$ exprimant une contrainte sur les termes $\mathbf{W}$ et/ou $\mathbf{H}$. Les formules de mise à jour multiplicative sont alors obtenues en faisant apparaitre au numérateur et au dénominateur de l'équation sans contrainte (2) les termes respectivement négatifs et positifs du gradient des contraintes additionnelles. On gardera toutefois à l'esprit qu'aucune preuve de convergence n'est donnée dans le cas général sous contrainte.

1.1.2 Application à la séparation de sources semi-supervisée

Dans un contexte réaliste, la séparation de sources audio consiste à extraire un signal cible - pour lequel on dispose d'informations a priori - d'un fond sonore inconnu à l'avance. Dans notre cas, la source d'intérêt est la voix d'un locuteur. Le spectrogramme du mélange peut être alors vu comme la somme des termes correspondant à la voix et des termes correspondant au bruit, ce qu'on écrit classiquement :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{W}_{ref}\mathbf{H} + \mathbf{W}^{\mathbf{N}}\mathbf{H}^{\mathbf{N}} \tag{3}$$

Dans le cas semi-supervisé dans lequel on se place ici, les bases d'un locuteur $\mathbf{W}_{ref}$ sont apprises à l'avance sur un signal non bruité et fixées pour la séparation, tandis que les bases du fond sonore $\mathbf{W}^{\mathbf{N}}$ sont laissées libres au cours de la mise à jour. Le minimum de $d_\beta(\mathbf{X}|\mathbf{V})$ plus d'éventuelles autres fonctions de coût est alors obtenu par une descente de gradient par rapport à toutes les matrices laissées libres ($\mathbf{H}$, $\mathbf{W}^{\mathbf{N}}$ et $\mathbf{H}^{\mathbf{N}}$).

Après convergence de la NMF, le spectrogramme de la partie correspondant à la voix est alors reconstitué par le produit $\mathbf{W}_{ref}\mathbf{H}$. Le signal sonore correspondant est alors restitué par exemple par un filtrage de Wiener [Benaroya and Bimbot, 2003], [Le Roux and Vincent, 2013] :

$$\mathbf{X} = \frac{\mathbf{W}_{ref}\mathbf{H}}{\mathbf{W}_{ref}\mathbf{H} + \mathbf{W}^{\mathbf{N}}\mathbf{H}^{\mathbf{N}}}$$

Cependant, la factorisation NMF obtenue en minimisant simplement la β-divergence $d_\beta(\mathbf{X}|\mathbf{V})$ d'un signal bruité n'a aucune raison de correspondre à une solution qui soit effectivement une séparation de la source cible et du bruit. En effet, on peut écrire $\mathbf{V}$ d'une infinité de façons dans lesquelles la répartition se fera différemment entre les termes correspondant à la source cible et les termes correspondants au bruit. En d'autres termes, des composantes bruitées peuvent se retrouver dans celles de la parole, et vice versa.

Il est donc nécessaire d'introduire des hypothèses supplémentaires sur le modèle de signal utilisé de façon à contraindre la convergence de la NMF vers une solution satisfaisante. Ces hypothèses peuvent s'exprimer sous forme de contraintes sur les matrices du modèle de signal, ou par une nouvelle formulation du modèle lui-même. Dans le cas de la voix et de nombreux instruments, c'est cette deuxième approche qui sera privilégiée, avec l'introduction dans le formalisme de la NMF du modèle source/filtre [Durrieu et al., 2009].

1.2 Séparation semi-supervisée de la parole

1.2.1 Rappels sur le modèle source-filtre

Un signal de parole peut être modélisé comme le résultat du filtrage convolutif par un système résonateur d'un signal produit par un système excitateur. Cette excitation peut-être un signal périodique dans le cas des phonèmes voisés, ou un signal non-périodique dans le cas des phonèmes bruités.

Dans le domaine fréquentiel, cette convolution s'exprime sous forme d'un produit. Un phonème est ainsi le résultat de la multiplication d'une excitation (composée par exemple des harmoniques d'une certaine fondamentale ou d'un signal de bruit) par un filtre (l'enveloppe spectrale du phonème) comme représenté sur la figure 2 :

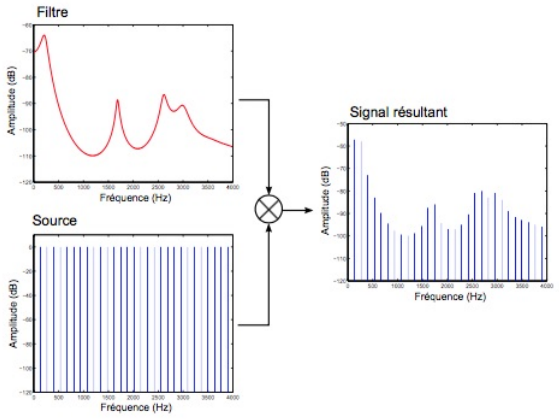


FIGURE 2 – Modèle source/filtre : filtrage convolutif d'un système excitateur

Les excitations et le filtre varient conjointement au cours du temps et ces variations correspondent à la succession de phonèmes qui forment la parole [Durrieu et al., 2009], visualisée sur la figure 3 :

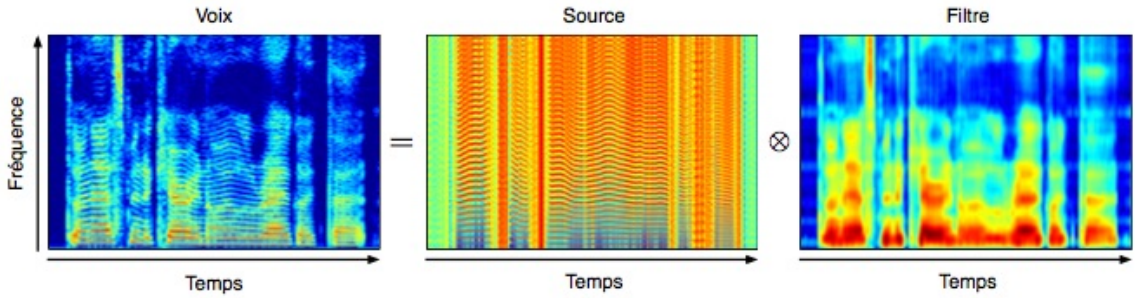


FIGURE 3 – Spectrogrammes du modèle source/filtre d'un signal de voix et de ses composantes

1.2.2 Formulation du modèle source/filtre dans un cadre NMF

La matrice du spectrogramme d'un signal de parole peut donc être représentée comme un produit de Hadamard de deux matrices représentant respectivement la source et le filtre. Chacune de ces deux matrices peut être factorisée comme le produit d'une matrice de bases spectrales et d'une matrice d'activation, ce qui s'écrit sous la forme suivante :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{W^{ex}H^{ex}} \otimes \mathbf{W^{\Phi}H^{\Phi}} + \mathbf{W^{N}H^{N}}$$

L'intérêt du modèle source/filtre est de pouvoir doter chaque matrice de la factorisation NMF d'une signification physique :

— $\mathbf{W^{ex}}$ est la matrice $F \times K^{ex}$ des bases spectrales de la source du modèle source/filtre, qui sont des peignes harmoniques correspondant aux K^{ex} fondamentales du modèle de la source, plus un certain nombre de bases constituées de bruit blanc permettant de rendre compte des phonèmes non-voisés. Elle est donc laissée *fixe* lors de la NMF.
— $\mathbf{H^{ex}}$ est la matrice $K^{ex} \times T$ des activations des bases de la source du modèle source/filtre. Elle est donc laissée *libre* lors de la NMF.
— $\mathbf{W^{\Phi}}$ est la matrice $F \times K^{\Phi}$ des bases spectrales du filtre, qui correspondent aux enveloppes spectrales des différents phonèmes. **L'apprentissage dans un cadre semi-supervisé est fait sur ces bases dans un contexte non bruité, et ces bases apprises sont alors utilisées pour effectuer une séparation dans un contexte bruité**. Cette matrice est donc laissée *fixe* lors de la NMF.
En pratique, l'apprentissage des formants des différents phonèmes pour un locuteur donné est réalisé en utilisant l'algorithme *True enveloppe* sur un corpus de phrases connues [Villavicencio et al., 2006], [Bouvier, 2015].
— $\mathbf{H^{\Phi}}$ est la matrice $K^{\Phi} \times T$ des activations des bases du filtre du modèle source/filtre. Elle est donc laissée *libre* lors de la NMF.

L'ensemble de cette décomposition peut alors se visualiser ainsi :

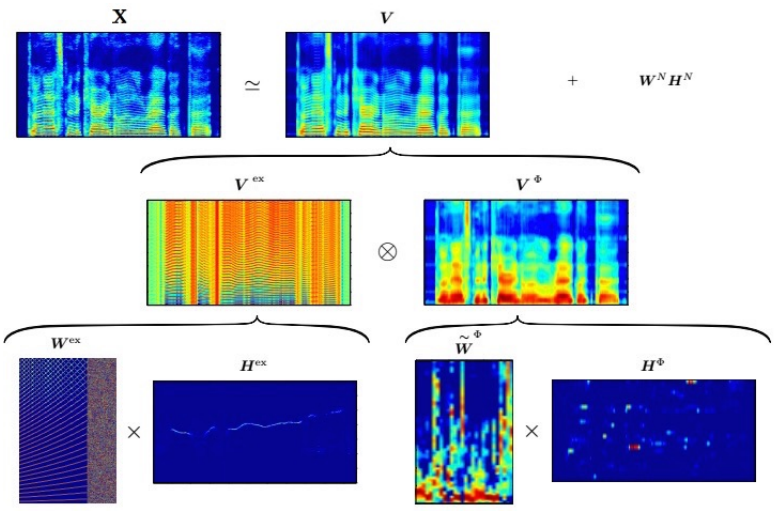


FIGURE 4 – Décomposition NMF d'un spectrogramme d'un modèle source/filtre d'un signal de voix

La compréhension de la signification physique permet ainsi de faire des hypothèses importantes sur les différentes contraintes que l'on peut imposer aux différentes matrices au cours de leur mise à jour, e.g. des

contraintes de parcimonie sur $\mathbf{H}^{\mathbf{ex}}$ et $\mathbf{H}^{\Phi}$ pour rendre compte de l'hypothèse qu'une voix parlée ne prononce qu'un seul phonème à la fois [Bouvier, 2015].

Le coût total à minimiser est alors la somme de $d_\beta(\mathbf{X}|\mathbf{V})$ et d'une fonction de coût exprimant les différentes contraintes additionnelles (parcimonie, etc.). La séparation proprement dite est alors obtenue en estimant les différentes matrices laissées libres, i.e. $\mathbf{H}^{\mathbf{ex}}$, $\mathbf{H}^{\Phi}$, $\mathbf{W}^{\mathbf{N}}$ et $\mathbf{H}^{\mathbf{N}}$, minimisant cette fonction de coût total.

1.2.3 Positionnement du problème

L'introduction du modèle source/filtre et la formalisation des différentes hypothèses (parcimonie, continuité temporelle) qui décrivent au mieux la réalité physique dans le cadre de ce modèle permettent de contraindre efficacement la convergence de la NMF au cours de la séparation. L'algorithme basé sur le modèle source/filtre permet en effet d'obtenir - après un paramétrage adéquat des différentes contraintes retenues - des résultats meilleurs que l'état de l'art pour un scénario semi-supervisé, i.e. avec apprentissage préalable des bases sur un locuteur de référence [Bouvier, 2015].

On s'intéresse à la comparaison des résultats obtenus pour les deux scenarii suivants :

— la séparation est effectuée sur un locuteur cible à partir des bases $\mathbf{W}^{\Phi}{}_{ref}$ apprises au préalable sur deux autres phrases non bruitées prononcées par ce *même* locuteur (noté ci-dessous Cible vs. Cible)
— la séparation est effectuée sur un locuteur cible à partir des bases $\mathbf{W}^{\Phi}{}_{ref}$ apprises au préalable sur deux autres phrases non bruitées prononcées par *un autre* locuteur (noté ci-dessous Cible vs. autre Référence)

Nous verrons en détails dans le chapitre 3 le protocole et le corpus utilisés pour effectuer les mesures de séparation, ainsi que les indicateurs retenus. Pour l'instant, retenons simplement que, pour un même corpus de différentes phrases mélangées à différents bruits avec trois rapports signal/bruit (SNR) de -6dB, 0dB et +6dB, on obtient les résultats suivants pour une séparation par NMF avec le modèle source/filtre décrit ci-dessus et les contraintes idoines décrites dans [Bouvier, 2015] :

	SDR			PESQ		
SNR (dB)	-6 dB	0 dB	+6 dB	-6 dB	0 dB	+6 dB
Cible vs. Cible	5.6	9.8	12.5	2.0	2.3	2.6
Cible vs. autre Référence	4.8	8.7	10.8	1.9	2.2	2.5

TABLE 1 – Résultats obtenus pour une séparation à partir du modèle source/filtre

On observe qualitativement que les résultats sont sensiblement moins bons dès lors que la séparation a lieu sur un locuteur cible qui n'est pas celui pour lequel l'apprentissage a eu lieu, ce qui avait déjà été souligné par [Sun and Mysore, 2013]. Ils sont même d'autant moins bons que le locuteur cible est d'un autre genre que le locuteur de référence (on obtient pour un SNR de 0dB un SDR de 8.5 si la cible est une femme et la référence est un homme et de 8.1 dans le cas contraire).

Ces résultats ne sont pas surprenants, puisque les enveloppes spectrales pour un même phonème présentent des différences d'un locuteur à l'autre, et *a fortiori* entre des locuteurs de sexe différent, comme cela apparait clairement sur la figure 5 :

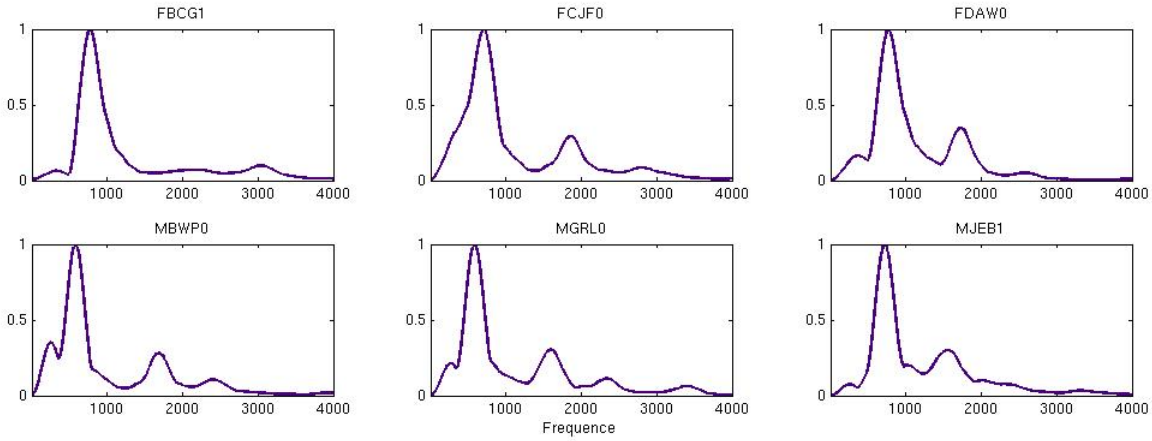


FIGURE 5 – Les enveloppes spectrales du phonème /a/ pour 3 locuteurs femme (en haut, référencées FBCO1, FCJF0 et FDAW0) et 3 locuteur hommes (en bas, référencés MBWP0, MGRL0 et MJEB1) issus de la base TIMIT

Le fait que les bases $\mathbf{W}^{\Phi}$ utilisées pour la séparation ne soient pas celles du locuteur cible et qu'elles soient maintenues *fixes* au cours de la séparation ne peut en effet pas permettre d'obtenir des résultas similaires à ceux qui sont obtenus lorsque les bases apprises $\mathbf{W}^{\Phi}$ sont celles du locteur cible.

Mais à l'inverse, si on laisse les bases $\mathbf{W}^{\Phi}$ *libres*, tout l'intérêt de l'apprentissage disparait : en l'absence de toute autre contrainte, nous n'avons en effet plus aucune garantie que la répartition de l'énergie du signal se fasse correctement entre les bases de voix et les bases de bruit. De fait, on obtient des résultats sensiblement moins bons en laissant $\mathbf{W}^{\Phi}$ libre sans contrainte particulière au cours de la séparation. Par exemple, pour un SNR = 0 dB, on obtient :

	SDR		PESQ	
	$\mathbf{W}_{ref}$ fixée	$\mathbf{W}_{ref}$ libre	$\mathbf{W}_{ref}$ fixée	$\mathbf{W}_{ref}$ libre
Cible vs. Cible	9.8	5.8	2.3	2.1
Cible vs. autre Référence	8.7	5.7	2.2	2.1

TABLE 2 – Comparaisons des résultats obtenus pour une séparation d'un mélange parole/bruit avec SNR=0 dB, à partir du modèle source/filtre avec $\mathbf{W}_{ref}$ fixée et $\mathbf{W}_{ref}$ libre

Ces observations nous amènent à poser le problème dans les termes suivants :

Dans le cas d'une séparation pour laquelle on ne dispose pas des bases $\mathbf{W}^{\Phi}$ apprises sur le locuteur cible, conserver $\mathbf{W}^{\Phi}$ fixée lors de la séparation ne permet pas aux bases de s'adapter (et donne des résultats moins bons que dans le cas où le locuteur cible est le locuteur de référence).

A l'inverse, laisser $\mathbf{W}^{\Phi}$ libre ne permet pas de contrôler sa mise à jour, et n'offre alors aucune garantie sur la façon dont la répartition de l'énergie du signal se fera entre les termes correspondant à la parole et ceux correspondant au bruit.

Comment est-il alors possible de permettre aux bases de s'adapter à un nouveau locuteur *au cours* de la NMF, tout en gardant un *contrôle* sur cette adaptation afin qu'elle se fasse de manière *cohérente* ?

Nous allons voir à présent que des solutions à ce problème ont déjà été envisagées dans le cadre de la séparation de sources audio, avec l'introduction, dans la formulation classique de la NMF, d'une matrice de déformation $\mathbf{D}$ sur les bases spectrales $\mathbf{W}^{\Phi}$.

1.3 Déformation des bases spectrales

1.3.1 Le modèle de déformation additive

Un modèle de déformation additive a été proposé par [Kitamura et al., 2013] dans un cadre de séparation par NMF semi-supervisée, i.e. dans lequel on dispose d'un apprentissage préalable des bases $\mathbf{W}$.

L'idée directrice consiste à dire que les bases apprises (sur un locuteur ou un instrument particulier) ne correspondent pas exactement à celles de la source cible, et que l'on peut modéliser cette différence par une matrice $\mathbf{G}$ s'ajoutant à la matrice de référence $\mathbf{W}_{ref}$. L'équation du problème s'écrit alors (les auteurs de cet article n'utilisant pas le modèle source/filtre) :

$$\mathbf{X} \simeq \mathbf{V} = (\mathbf{W_{ref}} + \mathbf{G})\mathbf{H} + \mathbf{W^N}\mathbf{H^N}$$

Dans ce modèle, la matrice cible s'écrit donc $\mathbf{W}_{cible} = \mathbf{W}_{ref} + \mathbf{G}$, où $\mathbf{W_{ref}}$ (en **violet**) est apprise au préalable et fixée une fois pour toutes. Les autres matrices en **noir** sont laissées libres.

Si la contrainte de non-négativité s'impose à $\mathbf{W}_{cible}$, ce n'est pas nécessairement le cas pour $\mathbf{G}$ qui peut prendre des valeurs négatives. Pour imposer la contrainte de positivité à la matrice $\mathbf{W}_{ref} + \mathbf{G}$ et limiter l'amplitude des éventuelles valeurs négatives de la déformation, les auteurs imposent ainsi aux coefficients de $\mathbf{G}$ la contrainte $\eta\mathbf{W}_{ij} + \mathbf{G}_{ij} \geq 0, \forall i, j \in [0...F] \times [0...K]$, avec $0 \leq \eta \leq 1$.

Mais en l'absence de toute autre hypothèse, il n'y a pas de raison particulière que la solution vers laquelle va converger la NMF offre une répartition correcte des bases spectrales de la cible entre celles de la déformation et celles du bruit.

Pour s'assurer que la séparation entre ces termes a effectivement lieu, les auteurs proposent donc d'introduire une contrainte de décorrélation entre les bases de chaque matrice spectrales $\mathbf{W}$, $\mathbf{G}$ et $\mathbf{W^N}$.

La fonction de coût global à minimiser s'écrit alors :

$$\mathscr{C}_{total} = d_\beta(\mathbf{V}|\mathbf{X}) + \|\mathbf{W}^T\mathbf{G}\|_F + \|\mathbf{W}^T\mathbf{W^N}\|_F + \|\mathbf{G}^T\mathbf{W^N}\|_F$$

1.3.2 Le modèle de déformation multiplicative

Un modèle de déformation multiplicative a été introduit par [Souviraa-Labastie et al., 2015] pour la séparation d'une source cible pour laquelle on dispose par ailleurs d'une source de référence : il peut par exemple s'agir de séparer une voix d'un fond musical dans une bande-son de film pour laquelle la musique est disponible seule, ou encore de séparer voix et musique sur un morceau pour lequel il existe une autre version (cover) dont on dispose de la piste de la voix seule.

Le principe général est de décrire conjointement la source cible et la (ou les) source(s) de référence sous la forme d'un système reliant les équations correspondant à chacune des sources considérées. Les auteurs envisagent ainsi dans un des exemples qu'ils proposent la séparation d'une voix sur un fond musical à l'aide de la même phrase prononcée par un autre locuteur en l'absence de bruit et d'une version originale de la musique seule.

En utilisant les indices vs, ms, vm, mm pour les termes correspondant respectivement à la voix seule, à la musique seule, à la voix dans le mix et à la musique dans le mix, le système des équations qui décrivent les différents signaux s'écrit alors :

$$\begin{cases} \mathbf{V_{mix}} &= \mathbf{W^{ex}_{vm}H^{ex}_{vm}} \otimes \mathbf{W^{\Phi}_{vm}H^{\Phi}_{vm}} \quad + \quad \mathbf{W^{ex}_{mm}H^{ex}_{mm}} \otimes \mathbf{W^{\Phi}_{mm}H^{\Phi}_{mm}} \quad + \quad \mathbf{W^{N}_{mix}H^{N}_{mix}} \\ \mathbf{V_{vs}} &= \mathbf{W^{ex}_{vs}H^{ex}_{vs}} \otimes (\mathbf{D^{f,\Phi}_{vs/vm}W^{\Phi}_{vm}D^{k,\Phi}_{vs/vm}H^{\Phi}_{vm}D^{t,\Phi}_{vs/vm}}) \\ \mathbf{V_{ms}} &= \mathbf{W^{ex}_{ms}H^{ex}_{mm}D^{t,ex}_{ms/mm}} \otimes \mathbf{W^{\Phi}_{mm}H^{\Phi}_{mm}} \quad + \quad \mathbf{W^{N}_{ms}H^{N}_{ms}} \end{cases}$$

où les matrices **W** et **H** correspondent classiquement aux bases spectrales et activations du modèle source/filtre.

Il est inutile de rentrer ici dans le détail de la méthode. Il nous suffit pour la suite de remarquer la présence des termes **D** qui sont les matrices correspondant à différentes déformations entre les bases et/ou les activations par une multiplication à droite ou à gauche des matrices **W** et **H**. Ce sont ces matrices de déformation que les auteurs proposent d'estimer au cours de la mise à jour NMF.

Par ailleurs, il nous sera utile pour la suite d'observer que dans ce système, les équations sont liées entre elles par différents termes qui leur sont communs. On note ainsi :

— en **violet** les termes qui sont gardés fixes tout au long de la séparation (ils correspondent aux bases des excitations qui sont des peignes harmoniques de différentes fondamentales et/ou un signal non-harmonique, et qui sont définies une fois pour toutes).
— en **carmin** les termes qui sont partagés par différentes sources "similaires" (la musique seule et la musique dans la bande-son d'une part, la parole seule d'un autre locuteur et la parole dans la bande-son d'autre part), et qui sont mis à jour *simultanément* pour les équations dans lesquelles ils apparaissent.
— en **noir** les termes laissés libres et qui sont mis à jour de manière classique.

La fonction de coût global à minimiser s'écrit alors comme la somme des β-divergences pour chaque source :

$$\mathscr{C}_{total} = d_\beta(\mathbf{V_{mix}}|\mathbf{X_{mix}}) + d_\beta(\mathbf{V_{vs}}|\mathbf{X_{vs}}) + d_\beta(\mathbf{V_{ms}}|\mathbf{X_{ms}})$$

1.3.3 Interprétation des différentes déformations

Les matrices de déformation modélisent des transformations différentes en fonction de la place qu'elles occupent dans les différentes équations et de la façon dont elles sont initialisées.

• Considérons par exemple le cas de la déformation multiplicative $\mathbf{W}_{cible} = \mathbf{D}\mathbf{W}_{ref}$. **D** est alors une matrice carrée $F \times F$, et chaque terme de la matrice cible déformée s'écrit en fonction des termes de la matrice de référence :

$$\mathbf{W}_{cible,fb} = \sum_k \mathbf{D}_{fk}\mathbf{W}_{ref,kb} \qquad \forall f, b \in \{1..F\} \times \{1..K\}$$

ce qu'on ne peut pas interpréter physiquement. Par contre, si la matrice de déformation n'a sur chaque ligne d'indice f qu'un seul terme non nul sur la colonne k_f, alors chaque terme de la matrice déformée s'écrit :

$$\mathbf{W}_{cible,fb} = \mathbf{D}_{fk_f}\mathbf{W}_{ref,k_fb}, \qquad \forall f, b \in \{1..F\} \times \{1..K\}$$

Si ce terme non nul est sur la diagonale ($k_f = f$), il s'agit alors d'un gain et la déformation revient à faire une égalisation. Si ce terme non nul est proche de la diagonale (eg. $k_f = f - 1$), alors la déformation revient à faire un décalage fréquentiel des composantes des bases (*pitch-shifting*).

On peut interpréter de la même façon la déformation multiplicative $\mathbf{H}_{cible} = \mathbf{H}_{ref}\mathbf{D}$, en observant que si la matrice de déformation n'a sur chaque colonne d'indice t qu'un seul ou quelques termes non nuls autour de la diagonale (eg. $k_t = \{t-1, .., t+1\}$), alors la déformation revient à faire un étalement ou un rétrécissement temporel des composantes des activations (*time-warping*).

• Considérons à présent la déformation additive $\mathbf{W}_{cible} = \mathbf{W}_{ref} + \mathbf{G}$. $\mathbf{G}$ est alors une matrice $F \times K$, et chaque terme de la matrice de bases spectrales déformée s'écrit en fonction des termes de la matrice de référence de la façon suivante :

$$\mathbf{W}_{cible,fb} = \mathbf{W}_{ref,fb} + \mathbf{G}_{fb} \qquad \forall f, b \in \{1..F\} \times \{1..K\}$$

A la différence des déformations multiplicatives décrites précédemment, chaque coefficient de la matrice de déformation n'a d'effet que sur le coefficient correspondant de la matrice de référence.

Soit alors une matrice $\mathbf{D}$ de mêmes dimensions que $\mathbf{G}$ telle que :

$$\mathbf{D}_{fb} = 1 + \frac{\mathbf{G}_{fb}}{\mathbf{W}_{ref,fb}} \qquad \forall f, b \in \{1..F\} \times \{1..K\}$$

La condition de non-négativité sur $\mathbf{W}_{ref} + \mathbf{G}$ implique alors que la matrice $\mathbf{D}$ est à coefficients positifs. On a par ailleurs :

$$\mathbf{W}_{cible} = \mathbf{W}_{ref} + \mathbf{G} = \mathbf{D} \otimes \mathbf{W}_{ref}$$

Il apparait ainsi que la déformation additive est équivalente à une déformation multiplicative terme à terme par une matrice non-négative que l'on peut interpréter comme l'expression d'un filtrage par un filtre de réponse en fréquence $\mathbf{D}$.

2 Contributions

Nous allons à présent montrer que les deux modèles de déformation abordés précédemment présentent certaines limitations pour le problème que nous avons posé au chapitre 1, ce qui nous amènera à proposer un autre modèle plus adapté. Nous décrirons alors différentes contraintes à imposer à ce modèle de déformation permettant de le contrôler et de s'assurer qu'il reste cohérent au cours de la séparation par NMF.

2.1 Intérêt et limites des modèles existants pour le problème posé

Nous allons montrer que le modèle de déformation additive de [Kitamura et al., 2013] est équivalent à un modèle multiplicatif terme à terme, d'un usage plus simple. Nous montrerons ensuite en quoi le modèle multiplicatif par matrice de [Souviraa-Labastie et al., 2015] ne peut pas convenir à notre problème, ce qui nous amènera alors à proposer un modèle multiplicatif par tenseur. Nous montrerons alors que ce modèle par tenseur peut également se ramener - sous certaines conditions - à un modèle multiplicatif terme à terme. Nous proposerons alors l'utilisation de ce modèle pour la résolution de notre problème.

2.1.1 Le modèle de déformation additif

Nous avons vu que le coût total à minimiser pour modèle de déformation additif est de la forme :

$$\mathscr{C}_{total} = d_\beta(\mathbf{V}|\mathbf{X}) + \mathscr{C}_{\mathbf{G}}(\mathbf{G}) + \mathscr{C} \quad \text{avec } \mathbf{V} = (\mathbf{W}+\mathbf{G})\mathbf{H} \tag{4}$$

où $\mathscr{C}_{\mathbf{G}}(\mathbf{G})$ et $\mathscr{C}$ correspondent aux composantes des contraintes respectivement dépendantes et indépendantes de $\mathbf{G}$.

Nous avons vu que $\mathbf{G}$ peut être à coefficients négatifs, ce qui empêche d'utiliser la règle de mise à jour multiplicative classique de la NMF, puisqu'on ne peut pas garantir que le pas de la descente de gradient soit toujours positif. C'est pourquoi les auteurs de [Kitamura et al., 2013] doivent introduire une fonction auxiliaire pour effectuer la mise à jour de $\mathbf{G}$.

Considérons alors l'écriture équivalente de $\mathbf{V}$ avec une déformation multiplicative terme à terme :

$$\mathbf{V} = (\mathbf{W}+\mathbf{G})\mathbf{H} = (\mathbf{D} \otimes \mathbf{G})\mathbf{H} \tag{5}$$

où $\mathbf{D}$ est telle que $\mathbf{D} = \mathbb{1} + \mathbf{G}./\mathbf{W}$, avec $\mathbb{1}$ étant la matrice dont tous les coefficients sont à 1 et où le symbole ./ représente une division terme à terme. Par construction, $\mathbf{D}$ est donc à coefficients positifs, puisque $\mathbf{W}+\mathbf{G}$ l'est.

On peut alors re-écrire la fonction de coût (4) uniquement en fonction de $\mathbf{D}$ (en remplaçant $\mathbf{G}$ par son expression en fonction de $\mathbf{D}$), et calculer alors le gradient de ce coût rapport à $\mathbf{D}$. Comme par construction $\mathbf{D}$ est à coefficients positifs, on peut alors utiliser la formule classique de mise à jour multiplicative dont le calcul est donné en annexe C et qui s'écrit pour l'itération n :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{-(n)}}{\mathbf{W} \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{+(n)}} \tag{6}$$

où $\mathbf{C}_{\mathbf{D}}^{-}$ et $\mathbf{C}_{\mathbf{D}}^{+}$ correspondent aux termes respectivement négatif et positif du gradient du coût de la contrainte imposée à $\mathbf{D}$.

On peut alors montrer que la mise à jour NMF de $\mathbf{D}$ est équivalente à la mise à jour de $\mathbf{G}$ (le détail des calculs est donné en annexe C). Cette équivalence s'écrit formellement :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{-(n)}}{\mathbf{W} \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{+(n)}} \iff \mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}_{\mathbf{G}}^{-(n)}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}_{\mathbf{G}}^{+(n)}} + \mathbf{\Gamma}^{(n)} \quad (7)$$

où $\mathbf{C}_{\mathbf{G}}^{-}$ et $\mathbf{C}_{\mathbf{G}}^{+}$ correspondent aux termes respectivement négatifs et positifs du gradient du coût de la contrainte imposée à $\mathbf{G}$, et où $\mathbf{\Gamma}^{(n)}$ est décrit ci-dessous.

Le terme de droite de l'équivalence ne dépend plus que de $\mathbf{G}$: il s'agit donc d'une formule possible de sa mise à jour multiplicative, qui ne nécessite pas d'hypothèse particulière sur le signe de ses coefficients. Le premier terme correspond à la règle de mise à jour classique d'une déformation additive que l'on obtiendrait dans le cas où $\mathbf{G}$ serait positive. Le deuxième terme $\mathbf{\Gamma}^{(n)}$ - dont l'expression analytique est donnée en annexe C - tend vers la matrice nulle en cas de convergence de la mise à jour, i.e. si $\mathbf{G}$ ne varie plus d'une itération à l'autre.

Cette équivalence montre que la mise à jour multiplicative de $\mathbf{G}$ converge si et seulement si la mise à jour multiplicative de $\mathbf{D}$ converge. Ce résultat est valable également en présence de contraintes sur $\mathbf{G}$ que l'on exprime comme autant de contraintes sur $\mathbf{D}$ selon les formules de passage décrites dans l'annexe C.

Nous proposons alors - puisque ces deux modèles, s'ils convergent, convergent à la même vitesse vers la même solution $\mathbf{W}+\mathbf{G} = \mathbf{D} \otimes \mathbf{W}$ - d'utiliser plutôt le modèle multiplicatif terme à terme qui a l'avantage d'être plus simple d'un point de vue calculatoire.

2.1.2 Le modèle de déformation multiplicatif par matrice

Déformation à droite

Considérons d'abord la déformation $\mathbf{W}_{cible} = \mathbf{W}_{ref}\mathbf{D}$, où $\mathbf{D}$ est alors une matrice $K \times K$.
Chaque terme de $\mathbf{W}_{cible}$ s'écrit $\mathbf{W}_{cible,fb} = \sum_k \mathbf{W}_{ref,fk}\mathbf{D}_{kb}$, ce qui signifie que chaque nouvelle b^{eme} base est un mélange de l'ensemble des bases de référence. Ce résultat n'est pas motivé d'un point de vue de l'interprétation physique (la re-estimation d'un phonème ne doit reposer que sur l'enveloppe spectrale disponible de ce phonème, et non pas être le résultat d'un mélange de phonèmes). Cette déformation ne peut donc pas convenir à notre problème.

Déformation à gauche

Considérons ensuite la déformation $\mathbf{W}_{cible} = \mathbf{D}\mathbf{W}_{ref}$, où $\mathbf{D}$ est alors une matrice $F \times F$.
Chaque terme de $\mathbf{W}_{cible}$ s'écrit $\mathbf{W}_{cible,fb} = \sum_k \mathbf{D}_{fk}\mathbf{W}_{ref,kb}$, ce qui signifie que chaque point en fréquence de la nouvelle base est une combinaison linéaire de toutes fréquences de la base de référence. Là encore, ce résultat n'est pas motivé d'un point de vue de l'interprétation physique (les déformations doivent être locales en fréquence, chaque formant de l'enveloppe spectrale se modifiant légèrement d'un locuteur à l'autre).

Nous avons vu que les auteurs de [Souviraa-Labastie et al., 2015] ont alors recours à des hypothèses fortes sur $\mathbf{D}$ pour pouvoir donner un sens physique à ce modèle de déformation (par exemple que $\mathbf{D}$ est diagonale ou n'a de coefficients non nuls qu'autour de la diagonale). Mais en faisant ce type d'hypothèse, on réduit la dimensionalité du problème : par exemple en supposant la matrice $\mathbf{D}$ diagonale, on passe d'un

problème initial sur-déterminé ($F \times F$ coefficients non nuls de $\mathbf{D}$ pour recomposer K bases $F \times 1$ avec $K < F$) à un problème sous-déterminé (F coefficients non nuls de $\mathbf{D}$ pour recomposer K bases $F \times 1$).

Or, l'observation de la variabilité des formants d'un phonème à l'autre entre deux locuteurs invalide l'idée qu'une seule et même déformation puisse rendre compte des différences entre toutes les bases à la fois :

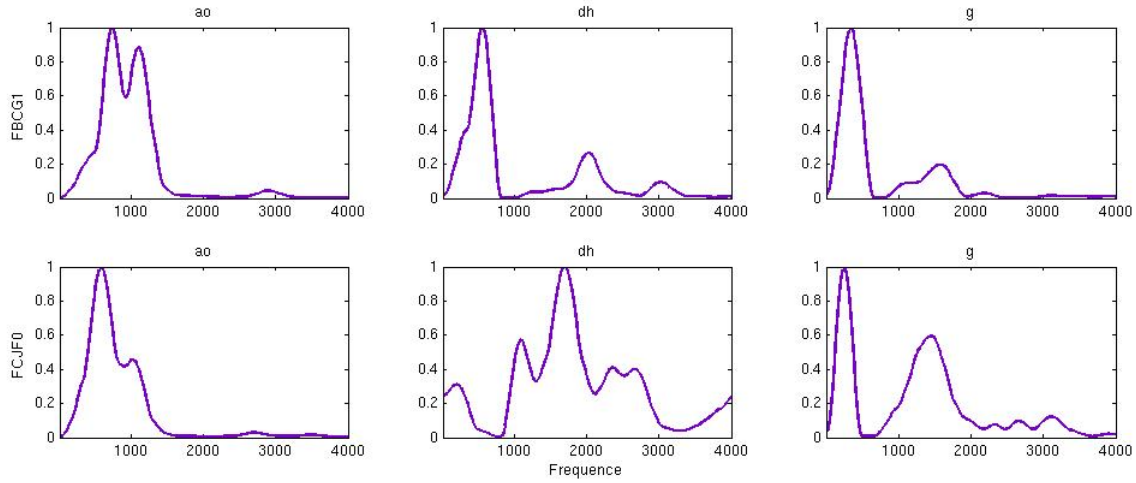


FIGURE 6 – Formants de trois phonèmes /ao/, /eh/, /g/ pour 2 locuteurs (FBCG1 en haut et FCJF0 en bas)

Comme on peut l'observer sur la figure 6, il n'y a aucune raison a priori pour que la déformation qui permette de passer d'un locuteur à l'autre pour le formant du phonème /ao/ soit la même déformation entre les formants du phonème /eh/ ou encore ceux du phonème /g/.

Le modèle multiplicatif avec une matrice commune à toutes les bases ne peut donc pas convenir à notre problème : soit il correspond à un problème sur-déterminé, et il est difficile à contraindre en l'absence d'interprétation physique, soit il correspond à un problème sous-déterminé qui n'a pas de raison particulière de donner des résultats cohérents. Nous allons donc proposer à présent un autre modèle pour pallier à cette difficulté.

2.2 Proposition d'un nouveau modèle de déformation

Pour s'assurer qu'une base du locuteur cible est bien obtenue à partir de l'adaptation d'une seule autre base correspondante du locuteur de référence, il faut renoncer à utiliser une unique matrice de déformation multiplicative commune à toutes les bases. Nous allons donc plutôt envisager dans un premier temps d'utiliser une matrice de déformation différente pour *chaque* base. Nous verrons alors que ce modèle constitue un modèle sur-déterminé qu'il est possible de simplifier sous certaines conditions.

2.2.1 Le modèle de déformation multiplicative par tenseur

Notons $\boldsymbol{\mathfrak{m}}_b$ et $\mathbf{w_b}$ la b^{eme} base respective du locuteur cible et du locuteur de référence, et écrivons que :

$$\boldsymbol{\mathfrak{m}}_b = \boldsymbol{\Delta_b} \mathbf{w_b} \quad \forall b \in \{1...K\} \tag{8}$$

où $\boldsymbol{\Delta_b}$ est une matrice $F \times F$. En notant $\boldsymbol{\Delta}$ le tenseur constitué des K matrices $\boldsymbol{\Delta_b}$, on convient d'écrire le modèle sous la forme $\mathbf{W}_{cible} = \boldsymbol{\Delta} \mathbf{W}_{ref}$.

Considérons la déformation d'une base en particulier (pour alléger les notations, on n'indiquera pas l'indice b dans ce qui suit) en supposant pour l'instant $\boldsymbol{\mathfrak{m}}$ et $\mathbf{w}$ connus, et cherchons $\boldsymbol{\Delta}$ qui est alors l'inconnue de l'équation (8). Il s'agit d'un problème sur-déterminé (puisque $\boldsymbol{\Delta}$ a $F \times F$ coefficients alors que $\boldsymbol{\mathfrak{m}}$ et $\mathbf{w}$ n'en ont que F chacun). Le déterminant d'un tel système est nul, il y a donc soit une infinité de solutions, soit aucune solution.

Or il existe une solution exacte à l'équation (8) : en effet, si l'on note $\mathbf{w}^\dagger$ le pseudo-inverse de $\mathbf{w}$, i.e. $\mathbf{w}^\dagger = \mathbf{w}^T(\mathbf{w}^T\mathbf{w})^{-1}$, alors $\boldsymbol{\Delta} = \boldsymbol{\mathfrak{m}}\mathbf{w}^\dagger$ est solution de (8) puisque $\boldsymbol{\mathfrak{m}}\mathbf{w}^\dagger\mathbf{w} = \boldsymbol{\mathfrak{m}}\mathbf{w}^T(\mathbf{w}^T\mathbf{w})^{-1}\mathbf{w} = \boldsymbol{\mathfrak{m}}\mathbf{w}^T\mathbf{w}(\mathbf{w}^T\mathbf{w})^{-1} = \boldsymbol{\mathfrak{m}}$, ($\mathbf{w}^T\mathbf{w}$ étant un scalaire).

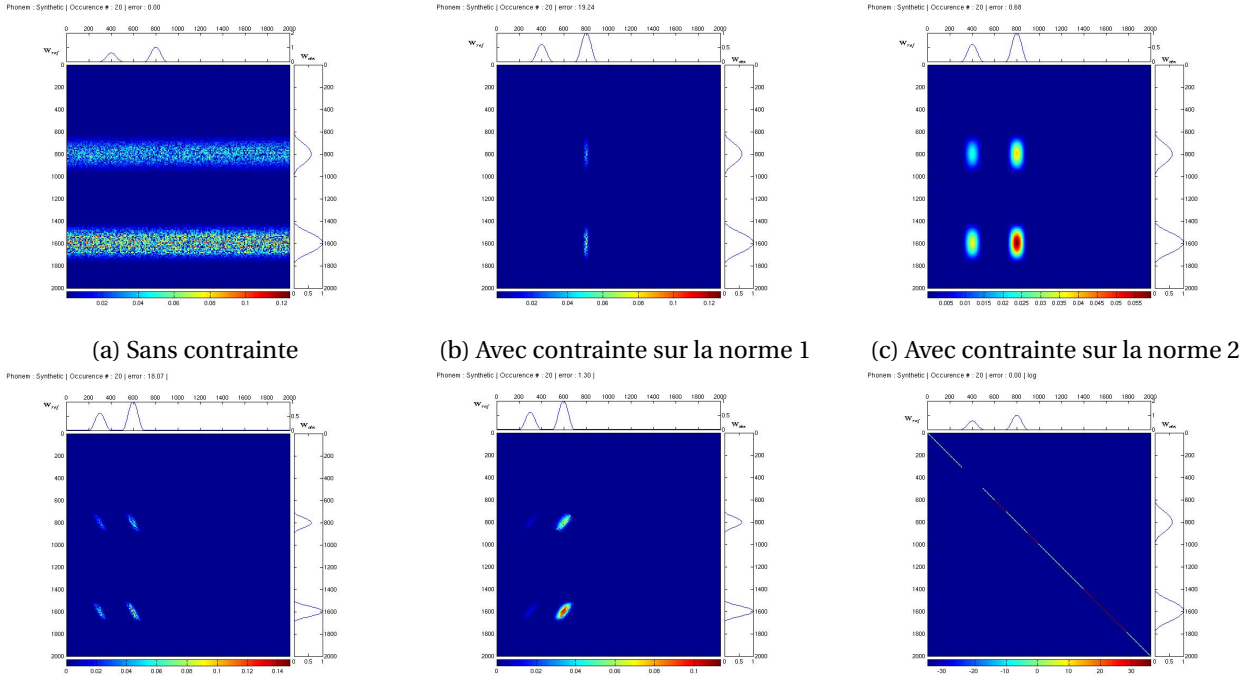
Il existe donc une infinité de matrices $\boldsymbol{\Delta}$ solutions de (8), i.e. qui permettent de transformer une base de référence en en une base cible donnée.

D'autres solutions de l'équation (8) peuvent être estimées en minimisant par rapport à $\boldsymbol{\Delta}$ une fonction de coût $\mathscr{C}_{total}$ telle que :

$$\mathscr{C}_{total}(\boldsymbol{\Delta}) = d_\beta(\boldsymbol{\mathfrak{m}}|\boldsymbol{\Delta}\mathbf{w}) + \mathscr{C}(\boldsymbol{\Delta})$$

où $d_\beta(\boldsymbol{\mathfrak{m}}|\boldsymbol{\Delta}\mathbf{w})$ est la β-divergence entre $\boldsymbol{\mathfrak{m}}$ et $\boldsymbol{\Delta}\mathbf{w}$, et $\mathscr{C}(\boldsymbol{\Delta})$ une contrainte que l'on impose à $\boldsymbol{\Delta}$. Chaque contrainte permet de réduire la sur-détermination du problème et d'aboutir à une solution différente.

La figure 7 représente par exemple différentes matrices $\boldsymbol{\Delta}$ solution de l'équation (8) pour différentes contraintes sur $\boldsymbol{\Delta}$, e.g. minimiser sa norme l_1 (fig. 7-b) ou l_2 (fig. 7-c), ou être diagonale (fig. 7-f). Dans cet exemple, $\boldsymbol{\mathfrak{m}}$ et $\mathbf{w}$ sont deux signaux synthétiques composés de 2 formants décalés en fréquence et $\boldsymbol{\Delta}$ est initialisée aléatoirement (le détail des calculs des gradients des contraintes est donné en annexe D) :



(a) Sans contrainte

(b) Avec contrainte sur la norme 1

(c) Avec contrainte sur la norme 2

(d) Distance l_1 avec une matrice de Toeplitz

(e) Distance l_2 avec une matrice de Toeplitz

(f) Forme diagonale

FIGURE 7 – Différentes solutions $\boldsymbol{\Delta}$ de l'équation (8) avec diverses contraintes (bases synthétiques)

Dans cette infinité de solutions, il en existe au moins deux qui ont une forme analytique :

— celle qui minimise la norme 2 de $\mathbf{\Delta}$ dans le cas où la β-divergence est la distance euclidienne est la solution analytique $\mathbf{\Delta} = \boldsymbol{\mathfrak{m}}\mathbf{w}^{\dagger}$ (cf. annexe D).
— celle où $\mathbf{\Delta}$ est diagonale et où son vecteur diagonale $F \times 1$ que l'on notera $\mathbf{d}$ est tel que $\mathbf{d} = \boldsymbol{\mathfrak{m}}./\mathbf{w}$ où ./ est une division terme à terme (puiqu'alors $\boldsymbol{\mathfrak{m}} = \mathbf{\Delta}\mathbf{w} = \mathbf{d} \otimes \mathbf{w}$)

La solution diagonale de (8) est particulièrement intéressante, parce que c'est une des solutions de la version déterminée du problème (F coefficients non nuls au lieu de $F \times F$). Soit alors $\mathbf{\Delta}$ une autre solution quelconque de (8), on peut écrire en toute généralité :

$$\boldsymbol{\mathfrak{m}} = \mathbf{\Delta}\mathbf{w} = \mathbf{d} \otimes \mathbf{w} \tag{9}$$

Si l'on revient à l'écriture tensorielle de la déformation des K bases, on écrira donc pour chaque b^{eme} base :

$$\boldsymbol{\mathfrak{m}}_b = \mathbf{\Delta}_b\mathbf{w}_b = \mathbf{d}_b \otimes \mathbf{w}_b \qquad \text{avec} \qquad \mathbf{d}_b = (\mathbf{\Delta}_b\mathbf{w}_b)./\mathbf{w}_b \qquad \forall b \in \{1...K\} \tag{10}$$

En notant alors $\mathbf{W}_b$ et $\mathbf{D}_b$ les matrices $F \times K$ dont toutes les colonnes sont nulles sauf la b^e que l'on pose égale respectivement à $\mathbf{w}_b$ et $\mathbf{d}_b$, on peut écrire :

$$\mathbf{W}_{cible} = \mathbf{\Delta}\mathbf{W}_{ref} = \sum_b \mathbf{\Delta}_b\mathbf{W}_b = \sum_b \mathbf{D}_b \otimes \mathbf{W}_b = \mathbf{D} \otimes \mathbf{W}_{ref} \tag{11}$$

Cela signifie que si l'on décrit la déformation entres les bases de deux locuteurs différents par un produit tensoriel, on peut toujours la décrire également comme une déformation par un produit de Hadamard matriciel. En d'autres termes, on peut toujours passer d'une solution particulière du cas sur-déterminé à une solution du cas déterminé qui lui est équivalente.

2.2.2 NMF des modèles de déformation tensorielle vs. multiplicative terme à terme

En pratique, chaque matrice $\mathbf{\Delta}_b$ du tenseur $\mathbf{\Delta}$ est estimée par NMF. La question qui se pose alors est donc la suivante : si l'on modélise la déformation par un tenseur $\mathbf{\Delta}$ soumis à des contraintes et que l'on estime la déformation correspondante par NMF, peut-on obtenir *la même* solution par la mise à jour d'une matrice déformation terme à terme $\mathbf{D}$ soumise à des contraintes correspondantes ?

On rappelle que la règle de mise à jour sous contraintes du tenseur $\mathbf{\Delta}$ à la n^e iteration s'écrit comme la succession des mises à jour sous contraintes de chacune de ses K matrices $\mathbf{\Delta}_b$ suivant la règle :

$$\mathbf{\Delta}_b^{(n+1)} = \mathbf{\Delta}_b^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)}}{\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)}} \qquad \forall b \in \{1...K\} \tag{12}$$

Les termes $\mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)}$ et $\mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)}$ sont deux matrices $F \times F$ qui correspondent respectivement à la partie négative et à la partie positive du gradient des contraintes sur chaque matrice $\mathbf{\Delta}_b$ à la n^e itération.

On peut alors montrer que la mise à jour de $\mathbf{\Delta}$ *implique* une mise à jour de $\mathbf{D}$ (le détail de la démonstration est donné en annexe E). Cette implication s'écrit formellement :

$$\boldsymbol{\Delta}_b^{(n+1)} = \boldsymbol{\Delta}_b^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}_{\boldsymbol{\Delta}_b}^{-\,(n)}}{\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}_{\boldsymbol{\Delta}_b}^{+\,(n)}} \quad \forall b \in \{1...K\} \Longrightarrow \mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{-(n)}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{+(n)}}$$

avec à chaque itération n l'égalité $\boldsymbol{\Delta}^{(n)}\mathbf{W} = \sum_b \boldsymbol{\Delta}_b^{(n)}\mathbf{W}_b = \mathbf{D}^{(n)} \otimes \mathbf{W}$.

A la différence du cas de la déformation additive, il n'y a pas ici d'équivalence entre les deux mises à jour, mais une simple implication. Ceci est du au fait que s'il est possible de trouver une matrice $\mathbf{D}$ vérifiant l'équation (11) à partir de l'expression de $\boldsymbol{\Delta}$ et de $\mathbf{W}$, l'inverse n'est pas vrai : on peut trouver une infinité de tenseurs satisfaisant (11) à partir de $\mathbf{D}$.

Les termes $\mathbf{C}_{\mathbf{D}}^{-(n)}$ et $\mathbf{C}_{\mathbf{D}}^{+(n)}$ sont deux matrices $F \times K$ qui correspondent respectivement à la partie négative et à la partie positive du gradient des contraintes $\mathscr{C}_{\mathbf{D}}$ que l'on impose alors sur la matrice $\mathbf{D}$ à la n^e itération. *Ces termes dépendent de* $\boldsymbol{\Delta}$, et l'on peut montrer (cf. annexe E) que :

$$\frac{\partial \mathscr{C}_{\boldsymbol{D}}}{\partial \boldsymbol{D}} = \sum_b \frac{\partial \mathscr{C}_{\boldsymbol{\Delta}_b}}{\partial \boldsymbol{\Delta}_b} \tilde{\mathbf{W}}_b \tag{13}$$

avec $\tilde{\mathbf{W}}_b = \mathbf{W}_b / \|\mathbf{W}_b\|_2^2$. Cette relation reste valable à chaque itération de la mise à jour NMF.

Dans le cas particulier où $\boldsymbol{\Delta}$ est laissé sans contraintes, alors $\mathbf{D}$ aussi, et les deux modèles de déformations convergent vers la même solution $\boldsymbol{\Delta}\mathbf{W} = \mathbf{D} \otimes \mathbf{W}$. Ce résultat est intéressant, car il montre que la convergence ne dépend pas de la forme que l'on impose à chaque $\boldsymbol{\Delta}_b$. Pour modéliser une déformation localisée en fréquence, nous envisagions initialement de contraindre chaque $\boldsymbol{\Delta}_b$ à n'avoir des coefficients non nuls que sur une bande d'une certaine largeur autour de sa diagonale. Il s'avère que le résultat sera le même en leur imposant d'être simplement diagonales, ce qui, d'un point de vue computationnel, permet de remplacer K multiplications matricielles par une simple multiplication terme à terme.

Dans le cas général, l'expression des termes du gradient des contraintes sur $\mathbf{D}$ dépend de l'expression du gradient des contraintes sur $\boldsymbol{\Delta}$: en présence de contraintes sur le tenseur, il n'est donc en général pas possible de ramener le cas sur-déterminé au cas déterminé.

Cependant, si l'expression du gradient des contraintes ne fait intervenir que des termes en $\boldsymbol{\Delta}_b\mathbf{W}_b$, alors on peut remplacer ces termes par leur équivalent $\mathbf{D}_b \otimes \mathbf{W}_b$. Pour ces cas-là, la déformation multiplicative par tenseur sous contraintes peut être écrite sous la forme d'une déformation multiplicative terme à terme sous contraintes correspondantes, et leur mise à jour par NMF converge à la même vitesse vers la même solution $\boldsymbol{\Delta}\mathbf{W} = \mathbf{D} \otimes \mathbf{W}$. On peut dans ces cas là - comme pour le cas sans contrainte - s'affranchir du calcul tensoriel.

En pratique, il s'avère comme nous le verrons que les contraintes que nous envisagerons ont une expression en $\boldsymbol{\Delta}\mathbf{W} = \mathbf{D} \otimes \mathbf{W}$ qui se retrouve dans l'expression de leur gradient (on en donne une illustration en annexe E et au chapitre 2.3).

Nous avons donc fait le choix d'utiliser le modèle de déformation multiplicatif terme à terme, parce qu'il permet de différencier la déformation de chacune des bases tout en étant plus simple à manipuler que le modèle de déformation par tenseur.

2.3 Application du modèle de déformation à la séparation de sources

L'équation du problème que l'on pose à présent s'écrit avec le modèle de déformation retenu :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{W}^{\mathbf{ex}}\mathbf{H}^{\mathbf{ex}} \otimes (\mathbf{D} \otimes \mathbf{W}^{\Phi})\mathbf{H}^{\Phi} + \mathbf{W}^{\mathbf{N}}\mathbf{H}^{\mathbf{N}} \tag{14}$$

où $\mathbf{D}$ est une matrice $F \times K^{\Phi}$ de déformation multiplicative terme à terme de la matrice des bases du filtre $\mathbf{W}^{\Phi}$ qui est apprise sur un locuteur de référence. Les autres matrices libres et mises à jour sont donc $\mathbf{H}^{\mathbf{ex}}$, $\mathbf{H}^{\Phi}$, $\mathbf{W}^{\mathbf{N}}$ et $\mathbf{H}^{\mathbf{N}}$.

En l'absence de contraintes, la déformation est laissée libre, ce qui revient à laisser libres les bases du filtre $\mathbf{W}^{\Phi}$. Nous avons vu au chapitre 1 qu'il n'y alors aucune raison pour que la répartition de l'énergie du signal se fasse correctement entre les bases déformées de voix et les bases de bruit. Nous allons donc étudier à présent les contraintes à imposer à notre modèle pour s'assurer que la déformation reste cohérente et que la séparation de la parole et du bruit a effectivement lieu.

2.3.1 Les contraintes de "petite déformation"

Dans un premier temps, nous avons envisagé des contraintes basées sur l'hypothèse selon laquelle les enveloppes spectrales d'un même phonème devaient présenter une forme relativement similaire d'un locuteur à l'autre - par exemple, un ou deux formants principaux dans des bandes de fréquences proches. Nous avons modélisé ce comportement par trois contraintes décrites ci-dessous que l'on qualifiera globalement de contraintes de "petite déformation".

Contrainte de "lissage spectral"

Cette contrainte, inspirée par [Virtanen, 2003], repose sur l'hypothèse selon laquelle chaque *bin* fréquentiel de l'enveloppe spectrale est proche de celui qui le précède et de celui qui le suit, i.e que l'enveloppe déformée reste lisse. Mathématiquement, cela se traduit par une dérivée continue en chaque point, ce qu'on traduit par une contrainte visant à minimiser l'écart entre chaque *bin* pour chaque base :

$$\mathscr{C}(\mathbf{D}) = \sum_{ij} |(\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j}|^2 \tag{15}$$

La mise à jour multiplicative de $\mathbf{D}$ s'écrit alors à partir du calcul du gradient de $\mathscr{C}(\mathbf{D})$ dont le détail est donné en annexe B :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + 2(\mathbf{D}^{(n)} \otimes \mathbf{W})^{\downarrow} + 2(\mathbf{D}^{(n)} \otimes \mathbf{W})^{\uparrow}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + 4\mathbf{D}^{(n)} \otimes \mathbf{W}} \tag{16}$$

où pour $\forall p, q \in \{1...F\} \times \{1...K\}$, $(\mathbf{D} \otimes \mathbf{W})^{\downarrow}_{pq} = (\mathbf{D} \otimes \mathbf{W})_{p-1,q}$ et $(\mathbf{D} \otimes \mathbf{W})^{\uparrow}_{pq} = (\mathbf{D} \otimes \mathbf{W})_{p+1,q}$.

Cette contrainte s'avérera particulièrement importante dès lors que l'on initialise $\mathbf{D}$ aléatoirement.

Contrainte de "petit gain"

On fait l'hypothèse que chaque enveloppe spectrale d'un même phonème d'un locuteur à l'autre est le résultat de petits gains sur chaque point fréquentiel. On exprime cette hypothèse en imposant une contrainte qui vise à minimiser une norme entre l'enveloppe spectrale déformée et l'enveloppe spectrale initiale, ce qu'on écrit :

$$\mathscr{C}(\mathbf{D}) = \|\mathbf{D} \otimes \mathbf{W} - \mathbf{W}\|^2 = \sum_{i,j} |\mathbf{D} \otimes \mathbf{W} - \mathbf{W}|^2_{ij} \tag{17}$$

La mise à jour multiplicative de $\mathbf{D}$ s'écrit alors à partir du calcul du gradient de $\mathscr{C}(\mathbf{D})$ dont le détail est donné en annexe B :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + 2\mathbf{W}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + 2\mathbf{D}^{(n)} \otimes \mathbf{W}} \tag{18}$$

Contrainte de "corrélation maximale entre phonèmes correspondants"

On fait l'hypothèse que chaque enveloppe spectrale d'un même phonème présente de fortes similitudes entre deux locuteurs. On exprime cette hypothèse en imposant une contrainte qui vise à maximiser la corrélation entre l'enveloppe spectrale déformée et l'enveloppe spectrale initiale de chaque phonème, ce qu'on écrit :

$$\mathscr{C}(\mathbf{D}) = -\|(\mathbf{D} \otimes \mathbf{W})^T \mathbf{W}\|^2 = -\sum_{i,j}((\mathbf{D} \otimes \mathbf{W})^T \mathbf{W})^2_{ij} \tag{19}$$

La mise à jour multiplicative de $\mathbf{D}$ s'écrit alors à partir du calcul du gradient de $\mathscr{C}(\mathbf{D})$ dont le détail est donné en annexe B :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + 2\mathbf{W}\mathbf{W}^T(\mathbf{D} \otimes \mathbf{W})}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T} \tag{20}$$

Le détail des calculs du gradient de chacune de ces trois contraintes par rapport à $\mathbf{D}$ est donné en annexe B, et les résultats obtenus pour chacune de ces contraintes sont décrits au chapitre 3.

2.3.2 Les contraintes de "déformation cohérente"

Comme nous l'avons déjà observé, la variabilité des enveloppes spectrales d'un même phonème d'un locuteur à l'autre jette un doute sur la réalité physique des hypothèses qui sous-tendent les contraintes de "petite déformation" : les figures 5 et 6 des chapitres précédents montrent qualitativement que l'hypothèse de petit gain et l'hypothèse de corrélation maximale en particulier ne sont pas toujours vérifiées en pratique.

Il est donc apparu nécessaire d'introduire d'autres contraintes visant à s'assurer de la cohérence de la déformation.

Contrainte de décorrélation avec le bruit

Comme nous l'avons déjà vu, il est important de s'assurer au cours de la séparation que la répartition de l'énergie du signal entre les termes correspondant à la source cible et les termes correspondant au bruit se fasse correctement. Dans l'article de [Kitamura et al., 2013], le modèle de signal est celui de la factorisation classique $\mathbf{V} = \mathbf{WH} + \mathbf{W^N H^N}$. La contrainte introduite par ces auteurs vise à décorréler les bases de $\mathbf{W}$ de celles du bruit $\mathbf{W^N}$ pour s'assurer que les composantes spectrales du bruit ne se retrouvent pas dans celles de la parole.

Dans notre cas, le modèle source/filtre que nous utilisons et que décrit l'équation (14) présente deux matrices de bases $\mathbf{W^{ex}}$ et $\mathbf{W^{\Phi}}$ qu'il n'est pas possible de ramener à une écriture de la forme $\mathbf{WH}$.

Décorréler les bases du bruit de celles des excitations n'a pas de signification particulière (cela supposerait qu'on ne peut ne retrouver aucun son harmonique dans le bruit, ce qui n'a aucune raison d'advenir dans le cas général). Décorréler les bases du bruit de celles du filtre n'a pas de signification particulière non plus : on ne s'attend de toutes façon pas à ce que les bases du filtre puissent être comparable à du bruit.

On a implémenté malgré tout cette contrainte de décorrélation qui vise à minimiser la corrélation entre chaque colonne de $\mathbf{D} \otimes \mathbf{W}^{\Phi}$ et chaque colonne de $\mathbf{W}^{\mathbf{N}}$, et qui s'écrit :

$$\mathscr{C}(\mathbf{D}, \mathbf{W}^{\mathbf{N}}) = \| (\mathbf{D} \otimes \mathbf{W}^{\Phi})^T \mathbf{W}^{\mathbf{N}} \|_2^2 \tag{21}$$

La mise à jour multiplicative de $\mathbf{D}$ s'écrit alors à partir du calcul du gradient de $\mathscr{C}(\mathbf{D})$ dont le détail est donné en annexe B :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + 2\mathbf{W}_N \mathbf{W}_N^T (\mathbf{D} \otimes \mathbf{W})} \tag{22}$$

Contrainte de réciprocité

L'idée est ici d'introduire une nouvelle équation faisant apparaître la matrice de déformation $\mathbf{D}$, de façon à disposer d'une information supplémentaire à laquelle la déformation doit se conformer pour rester cohérente.

Soit le locuteur cible inconnu, que l'on indicera par $\mathbf{0}$ par la suite, et un locuteur de référence connu, que l'on indicera par $\mathbf{r}$ par la suite. On note $\mathbf{V}_{0/r}$ l'estimation que l'on fait du signal observé $\mathbf{X}_0$ pour le locuteur cible à partir de l'apprentissage sur le locuteur de référence. L'idée est alors de se munir du signal non-bruité observé $\mathbf{X}_r$ sur lequel se fait l'apprentissage et conserver $\mathbf{H}_{\mathbf{r}}^{\mathbf{ex}}$, $\mathbf{W}_{\mathbf{r}}^{\Phi}$ et $\mathbf{H}_{\mathbf{r}}^{\Phi}$.

Jusqu'à présent, on se contente d'adapter avec une matrice $\mathbf{D}_{0/r}$ les bases du filtre du locuteur de référence pour estimer celles du locuteur cible (lors de la NMF visant à miniser $d_\beta(\mathbf{X}_0|\mathbf{V}_{0/r})$). On propose à présent d'adapter *conjointement* avec une matrice $\mathbf{D}_r$ les bases estimées obtenues pour le locuteur cible (lors d'une NMF visant cette fois à minimiser $d_\beta(\mathbf{X}_r|\mathbf{V}_r)$). On doit alors retrouver les bases du locuteur de référence connues, ce qui s'écrit $\mathbf{D}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r \simeq \mathbf{W}_r$. L'introduction de cette équation supplémentaire permet de guider la convergence vers une solution qui doit rester cohérente vis-à-vis de l'information disponible. Le système à résoudre s'écrit alors :

$$\begin{cases} \mathbf{V}_{0/r} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/r} \otimes \mathbf{W}_r^{\Phi})\mathbf{H}_0^{\Phi} + \mathbf{W}^N\mathbf{H}^N \\ \mathbf{V}_r & = \mathbf{W}^{ex}\mathbf{H}_r^{ex} \otimes (\mathbf{D}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r^{\Phi})\mathbf{H}_r^{\Phi} \end{cases} \tag{23}$$

avec ${\mathbf{W}^{\Phi}}_r$ et ${\mathbf{H}^{\Phi}}_r$ étant apprises et fixées , $\mathbf{D}_{0/r}$ étant partagée entre chaque paire $\mathbf{V}_{0/r}$ vs. $\mathbf{V}_r$, et $\mathbf{H}_{\mathbf{0}}^{\mathbf{ex}}$, $\mathbf{H}_{\mathbf{0}}^{\Phi}$, $\mathbf{W}^N$ et $\mathbf{H}^N$ étant laissées libres.

On cherche alors à estimer $\mathbf{D}_{0/r}$ et $\mathbf{D}_r$ minimisant :

$$\mathscr{C}(\mathbf{D}_{0/r}, \mathbf{D}_r) = d_\beta(\mathbf{X}_0|\mathbf{V}_{0/r}) + d_\beta(\mathbf{X}_r|\mathbf{V}_r) \tag{24}$$

Les mises à jour multiplicatives conjointes de $\mathbf{D}_{0/r}$ et $\mathbf{D}_r$ s'écrivent alors à partir du calcul du gradient de $\mathscr{C}(\mathbf{D}_{0/r}, \mathbf{D}_r)$ dont le détail est donné en annexe B :

$$\mathbf{D}_{0/r} = \mathbf{D}_{0/r} \otimes \frac{\mathbf{W}_r^{\Phi} \otimes ((\mathbf{X}_0 \otimes \mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{.(\beta-2)})\mathbf{H}_0^{\Phi T}) + (\mathbf{D}_r \otimes \mathbf{W}_r^{\Phi}) \otimes ((\mathbf{X}_r \otimes \mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{.(\beta-2)})\mathbf{H}_r^{\Phi T})}{\mathbf{W}_r^{\Phi} \otimes ((\mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{.(\beta-1)})\mathbf{H}_0^{\Phi T}) + (\mathbf{D}_r \otimes \mathbf{W}_r^{\Phi}) \otimes ((\mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{.(\beta-1)})\mathbf{H}_r^{\Phi T})} \tag{25}$$

$$\mathbf{D}_r = \mathbf{D}_r \otimes \frac{(\mathbf{D}_{0/r} \otimes \mathbf{W}_r^{\Phi}) \otimes ((\mathbf{X}_r \otimes \mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{.(\beta-2)})\mathbf{H}_r^{\Phi T})}{(\mathbf{D}_{0/r} \otimes \mathbf{W}_r^{\Phi}) \otimes ((\mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{.(\beta-1)})\mathbf{H}_r^{\Phi T})} \tag{26}$$

L'introduction du terme $d_\beta(\mathbf{X}_r|\mathbf{V}_r)$ dans la fonction de coût pemet de guider la convergence de $\mathbf{D}_{0/r}$ et $\mathbf{D}_r$ vers une solution pour laquelle $\mathbf{D}_{0/r} \otimes \mathbf{D}_r$ tend vers la matrice unité, ce qui est une information supplémentaire à respecter pour la convergence de $\mathbf{D}_{0/r}$ induite par le terme $d_\beta(\mathbf{X}_0|\mathbf{V}_{0/r})$.

Contrainte d'identité

L'idée est à nouveau d'introduire de nouvelles équations faisant intervenir $\mathbf{D}_{0/r}$ de façon à disposer d'une information de cohérence supplémentaire à respecter lors de la mise à jour.

Soit toujours le locuteur cible indicé par 0, et soient à présent R autres locuteurs de référence. On considère toujours un signal observé $\mathbf{X}_0$ pour le locuteur cible. Mais au lieu de faire une seule séparation à partir d'un seul locuteur de référence, on fait simultanément R séparations, une pour chaque référence, et l'on adapte pour chacune les bases apprises de référence. L'idée est que toutes ces déformations différentes de bases apprises sur des locuteurs différents doivent toutes converger vers la même matrice $\mathbf{W_0^{ex}}$ qui est celle du locuteur cible. Le système à résoudre s'écrit donc :

$$\begin{cases} \mathbf{V}_{0/1} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/1} \otimes \mathbf{W}_1^{\Phi})\mathbf{H}_0^{\Phi} + \mathbf{W}^N\mathbf{H}^N \\ & \dots \\ \mathbf{V}_{0/r} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/r} \otimes \mathbf{W}_r^{\Phi})\mathbf{H}_0^{\Phi} + \mathbf{W}^N\mathbf{H}^N \\ & \dots \\ \mathbf{V}_{0/R} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/R} \otimes \mathbf{W}_R^{\Phi})\mathbf{H}_0^{\Phi} + \mathbf{W}^N\mathbf{H}^N \end{cases}$$

où :

- $\mathbf{W}^{ex}$ est fixée une fois pour toutes
- chaque $\mathbf{H}_r^{ex}$, $\mathbf{H}_r^{\Phi}$ et $\mathbf{W}_r^{\Phi}$ est apprise et fixée, $\forall r \in \{0...R\}$
- chaque $\mathbf{D}_{0/r}$ est laissée libre
- $\mathbf{H}_0^{ex}$, $\mathbf{H}_0^{\Phi}$, $\mathbf{W}^N$ et $\mathbf{H}^N$ sont partagées par toutes les estimations $\mathbf{V}_{0/r}$ du même $\mathbf{X}_0$.

Pour chaque référence r, on cherche donc à faire converger $\mathbf{D}_{0/r} \otimes \mathbf{W_r^{\Phi}}$ vers la même matrice $\mathbf{\tilde{W}_0}$ des bases du locuteur cible. On introduit alors pour cela une contrainte sur le système que l'on appelera contrainte d'identité, et que l'on écrit

$$\mathscr{C}_{id}(\mathbf{D}_{0/r}) = \sum_{s=r}^{R} \|\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/s} \otimes \mathbf{W}_s\|_F^2 \qquad \text{pour } \forall r \in \{1...R\} \tag{27}$$

On cherche ensuite à estimer chaque $\mathbf{D}_{0/r}$ minimisant le coût total :

$$\mathscr{C} = \sum_{\rho=1}^{R} d_{\beta}(\mathbf{V}_{0/\rho}|\mathbf{X}_0) + \lambda_{id}\mathscr{C}_{id}(\mathbf{D}_{0/\rho})$$

où λ_{id} est un poids sur la contrainte d'identité que l'on peut faire varier au cours des tests.

L'écriture de ce système suppose que $\mathbf{H_0^{\Phi}}$ est partagée par les R équations. Dès lors, $\mathbf{D}_{0/r}$ et $\mathbf{W_r^{\Phi}}$ doivent avoir le même nombre de bases. En pratique, ce n'est pas toujours le cas, si par exemple certains phonèmes ne sont pas prononcés sur les phrases d'apprentissage. Imposer la contrainte d'identité requiert donc un traitement préalable des matrices $\mathbf{W_0^{\Phi}}$ pour faire un alignement et s'assurer que les bases manquantes sont remplies. Dans la pratique, on a choisi de prendre la base manquante chez une référence r dans la première autre référence r' pour laquelle la base en question est présente.

Les mises à jour multiplicatives conjointes de chaque $\mathbf{D}_{0/r}$ s'écrivent alors à partir du calcul du gradient de $\mathscr{C}$ dont le détail est donné en annexe B :

$$\mathbf{D}_{0/r} = \mathbf{D}_{0/r} \otimes \frac{\mathbf{W}_r^{\Phi} \otimes ((\mathbf{X}_0 \otimes \mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{\cdot(\beta-2)})\mathbf{H}_0^{\Phi T}) + 2\lambda_{id}\mathbf{W}_r \otimes \sum_{\rho=1}^{R} \mathbf{D}_{0/\rho} \otimes \mathbf{W}_{\rho}}{\mathbf{W}_r^{\Phi} \otimes ((\mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{\cdot(\beta-1)})\mathbf{H}_0^{\Phi T}) + 2\lambda_{id}R\mathbf{W}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r} \qquad \forall r \in \{1...R\}$$

Contrainte de réciprocité et d'identité simultanées

Nous avons alors envisagé d'imposer aux déformations de chaque $\mathbf{W}_\mathbf{r}^\Phi$ pour chaque référence les deux contraintes en même temps, i.e. à minimiser :

$$\mathscr{C} = \sum_{\rho=1}^{R} d_\beta(\mathbf{V}_{0/\rho}|\mathbf{X}_0) + (d_\beta(\mathbf{V}_\rho|\mathbf{X}_\rho) + \lambda_{id}\mathscr{C}_{id}(\mathbf{D}_{0/\rho})$$

pour résoudre le système suivant :

$$\begin{cases} \mathbf{V}_{0/1} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/1} \otimes \mathbf{W}_1^\Phi)\mathbf{H}_0^\Phi + \mathbf{W}^N\mathbf{H}^N \\ \mathbf{V}_1 & = \mathbf{W}^{ex}\mathbf{H}_1^{ex} \otimes (\mathbf{D}_1 \otimes \mathbf{D}_{0/1} \otimes \mathbf{W}_1^\Phi)\mathbf{H}_1^\Phi \\ & \dots \\ \mathbf{V}_{0/r} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/r} \otimes \mathbf{W}_r^\Phi)\mathbf{H}_0^\Phi + \mathbf{W}^N\mathbf{H}^N \\ \mathbf{V}_r & = \mathbf{W}^{ex}\mathbf{H}_r^{ex} \otimes (\mathbf{D}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r^\Phi)\mathbf{H}_r^\Phi \\ & \dots \\ \mathbf{V}_{0/R} & = \mathbf{W}^{ex}\mathbf{H}_0^{ex} \otimes (\mathbf{D}_{0/R} \otimes \mathbf{W}_R^\Phi)\mathbf{H}_0^\Phi + \mathbf{W}^N\mathbf{H}^N \\ \mathbf{V}_R & = \mathbf{W}^{ex}\mathbf{H}_R^{ex} \otimes (\mathbf{D}_R \otimes \mathbf{D}_{0/R} \otimes \mathbf{W}_R^\Phi)\mathbf{H}_R^\Phi \end{cases}$$

où :

— $\mathbf{W}^{ex}$ est fixée une fois pour toutes
— chaque $\mathbf{H}_r^{ex}$, $\mathbf{H}_r^\Phi$ et $\mathbf{W}_r^\Phi$ est apprise et fixée, $\forall r \in \{0...R\}$
— chaque $\mathbf{D}_{0/r}$ est partagée réciproquement par chaque couple $\mathbf{V}_{0/r}$ et $\mathbf{V}_r$, $\forall r \in \{0...R\}$
— chaque $\mathbf{D}_r$ est libre pour chaque $\mathbf{V}_r$, $\forall r \in \{0...R\}$
— $\mathbf{H}_0^{ex}$, $\mathbf{H}_0^\Phi$, $\mathbf{W}^N$ et $\mathbf{H}^N$ sont partagées par toutes les estimations $\mathbf{V}_{0/r}$ du même $\mathbf{X}_0$.

Les mises à jour multiplicatives conjointes de chaque $\mathbf{D}_{0/r}$ et chaque $\mathbf{D}_r$ s'écrivent alors à partir du calcul du gradient de $\mathscr{C}$ dont le détail est donné en annexe B :

$$\mathbf{D}_{0/r} = \mathbf{D}_{0/r} \otimes \frac{\mathbf{W}_r^\Phi \otimes ((\mathbf{X}_0 \otimes \mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{\cdot(\beta-2)})\mathbf{H}_0^{\Phi T}) + (\mathbf{D}_r \otimes \mathbf{W}_r^\Phi) \otimes ((\mathbf{X}_r \otimes \mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{\cdot(\beta-2)})\mathbf{H}_r^{\Phi T}) + 2\lambda_{id}\mathbf{W}_r \otimes \sum_{\rho=1}^{R} \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho}{\mathbf{W}_r^\Phi \otimes ((\mathbf{V}_0^{ex} \otimes \mathbf{V}_{0/r}^{\cdot(\beta-1)})\mathbf{H}_0^{\Phi T}) + (\mathbf{D}_r \otimes \mathbf{W}_r^\Phi) \otimes ((\mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{\cdot(\beta-1)})\mathbf{H}_r^{\Phi T}) + 2\lambda_{id}R\mathbf{W}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r} \tag{28}$$

$$\mathbf{D}_r = \mathbf{D}_r \otimes \frac{(\mathbf{D}_{0/r} \otimes \mathbf{W}_r^\Phi) \otimes ((\mathbf{X}_r \otimes \mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{\cdot(\beta-2)})\mathbf{H}_r^{\Phi T})}{(\mathbf{D}_{0/r} \otimes \mathbf{W}_r^\Phi) \otimes ((\mathbf{V}_r^{ex} \otimes \mathbf{V}_r^{\cdot(\beta-1)})\mathbf{H}_r^{\Phi T})} \tag{29}$$

pour $\forall r \in \{1...R\}$.

3 Résultats

Nous allons à présent présenter les résultats obtenus pour les différentes contraintes imposées au modèle de déformation décrits dans le chapitre précédent. Nous expliciterons tout d'abord le protocole et le corpus de signaux de voix utilisés pour les tests. Nous détaillerons ensuite les résultats obtenus pour chacun des scénarii envisagés. Nous terminerons enfin par une discussion des résultats.

3.1 Description du protocole de test

3.1.1 La base de test

Afin de pouvoir évaluer le modèle de déformation et les contraintes décrites au chapitre 2, d'une part, et de pouvoir comparer ces résultats aux résultats obtenus avec le modèle source/filtre sans déformation , d'autre part, nous avons utilisé la même base de tests que celle utilisée par [Bouvier, 2015].

Cette base a été construite à partir de la base de parole TIMIT [Zue et al., 1990] et de la base bruit QUT-NOISE [Dean et al., 2010].

TIMIT est une base de données de 3600 phrases en anglais-américain, prononcées par 360 locuteurs différents et enregistrées à 16kHz. Chaque fichier audio est accompagné d'un fichier texte correspondant comportant la segmentation de chaque phonème de référence prononcé dans la phrase. Chaque locuteur a enregistré 10 phrases, dont 2 sont identiques pour tous les locuteurs ; ces 2 phrases exposent toutes les variantes de prononciations (accents) des locuteurs.

20 locuteurs parmi les 360 ont été choisi aléatoirement pour notre base de test. L'apprentissage des bases $\mathbf{W}^{\Phi}$ pour chaque locuteur a été réalisé sur les 2 phrases communes à tous. Les tests proprement dit on été menés sur les 8 phrases restantes.

Ces 20 locuteurs, 10 hommes et 10 femmes ont pour identifiant TIMIT : FBCG1, FSDJ0, FCJF0, FDAW0, FDML0, FECD0, FETB0, FMBG0, FHEW0, FJXM0, MKLS0, MJEB1, MGRL0, MMRP0, MPGH0, MPDF0, MRAI0, MBWP0, MMGK0 et MTRR0.

QUT-NOISE contient 20 enregistrements de 30mn de différents environnements sonores (cuisine, rue, café, intérieur d'une voiture, etc) réalisés à 48kHz. Ils ont été sous-échantillonnés à 16kHz et des échantillons nécessaires à la création de la base de parole bruitée utilisée pour les tests ont été extraits aléatoirement.

Les 5 types de bruit utilisés sont les 4 issus de QUT-NOISE-TIMIT (nommés STREET-CITY-1, CAFE-CAFE-1, HOME-KITCHEN-1 et CAR-WINDOWNB-1), et du bruit blanc.

Chacune des 8 phrases des 20 locuteurs a alors été mélangée successivement avec les 5 types de bruit retenus pour 3 ratios signal-sur-bruit (SNR, pour Signal-to-Noise Ratio) différents : -6dB, +0dB et +3dB, soit 2400 phrases bruitées au total.

3.1.2 Les métriques utilisées

La qualité de la séparation de la parole et du bruit est mesurée à partir des indicateurs suivants :

• Le SDR, le SIR et le SAR définis dans [Vincent et al., 2006] :
— le SIR (*Signal-to-Interference Ratio*) quantifie, dans le signal reconstitué, le rapport en décibels entre le signal utile et les interférences venues des autres sources.
— le SAR (*Signal-to-Artifact Ratio*) quantifie, dans le signal reconstitué, le rapport en décibels entre le signal utile et les artefacts créés par l'algorithme.
— le SDR (*Signal-to-Distortion Ratio*) quantifie, dans le signal reconstitué, le rapport en décibels entre le signal utile et la somme des artefacts et des interférences ; il englobe le SIR et le SAR et caractérise la qualité de la séparation.

• Le PESQ (*Perceptual Evaluation of Speech Quality*) proposé dans [Rix et al., 2001] qui un critère MOS (*Mean Opinion Score*, ou Note d'opinion moyenne) qui caractérise la qualité de la restitution sonore du signal de voix reconstruit. Il varie entre 1 (très mauvais) et 5 (excellent, comparable à la version d'origine). Le PESQ a été développée originellement pour l'évaluation des réseaux téléphoniques et des codecs audio.

Les différents algorithmes et contraintes ont alors été exécutés sur les 800 phrases pour chaque SNR, et les résultats indiqués sont la moyenne des scores obtenus pour chaque phrase. Dans ce rapport, seuls le SDR et le PESQ ont été indiqués.

3.1.3 Le protocole de test utilisé

On rappelle à nouveau le modèle de mélange utilisé :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{W}^{\mathbf{ex}}\mathbf{H}^{\mathbf{ex}} \otimes (\mathbf{D} \otimes \mathbf{W}^{\Phi})\mathbf{H}^{\Phi} + \mathbf{W}^{\mathbf{N}}\mathbf{H}^{\mathbf{N}}$$

où :
— $\mathbf{W}^{\mathbf{ex}}$ est le dictionnaire d'excitations qui comprend 350 bases, divisées entre 250 bases périodiques (raies harmoniques plates) espacées tout les dixièmes de demi-tons entre 80 et 350Hz et 100 bases bruitées, construites à partir de bruit blanc.
— $\mathbf{W}^{\Phi}$ est la matrice des bases du filtre apprise pour un locuteur de référence. Selon les locuteurs, le nombre de bases (phonèmes) varie de 28 à 35 (certains phonèmes n'apparaissent pas dans les phrases prononcées par certains des locuteurs retenus).
— $\mathbf{W}^{\mathbf{N}}$ est la matrice des bases du bruit. Nous avons utilisé le nombre de bases qui permettait d'obtenir les meilleurs scores pour une séparation sans déformation, soit 25 bases de bruit [Bouvier, 2015].

R=1 locuteur de référence

On s'est d'abord placé dans le cas où l'apprentissage de $\mathbf{W}^{\Phi}$ se fait sur un seul locuteur référence (R=1). Chaque locuteur du corpus a été pris successivement comme locuteur cible, et on a effectué pour chaque locuteur cible une séparation en prenant successivement comme référence pour l'apprentissage de $\mathbf{W}^{\Phi}$ lui-même puis tous les autres locuteurs du corpus.

Chaque séparation s'est faite sur 7 des 8 phrases prononcées par chaque locuteur cible (la 8^e phrase du corpus a été reservée pour être utilisée comme signal observé $\mathbf{X}_r$ pour la contrainte de réciprocité).

On obtient alors un tableau de résultats 20 locuteurs cibles ×20 locuteurs de référence dont on extrait 4 moyennes :
— la moyenne des résultats pour les 20 cas où le locuteur cible est le locuteur de référence (noté ci-dessous Cible vs. Cible)

— la moyenne des résultats pour les 10×9 cas où le locuteur cible est une femme et le locuteur de référence est une femme différente du locuteur cible (Cible F vs. autre Référence F)
— la moyenne des résultats pour les 10×9 cas où le locuteur cible est un home et le locuteur de référence est un homme différent du locuteur cible (Cible H vs. autre Référence H)
— la moyenne des résultats pour les 20×19 cas où le locuteur cible est différent du locuteur cible sans hypothèse sur le genre (Cible vs. autre Référence)

R=2 à 5 locuteurs de référence

On s'est ensuite placé dans le cas de la contrainte d'identité où on réalise une séparation simultanée avec plusieurs locuteurs de référence.

Le nombre de combinaisons de locuteurs de référence est élevé. Pour R=2, il est de $C_{20}^2 = 190$, pour R=3 il est de $C_{20}^2 = 1140$, etc. Il n'a donc pas été possible de faire les tests sur toutes les combinaisons possibles. Nous avons donc limité le corpus à un locuteur cible et 7 locuteurs de référence (le locuteur cible lui-même, 3 locuteurs femme et 3 locuteurs hommes), ce qui faisait par exemple pour R=2 références parmi 7 = 21 combinaisons de 2 locuteurs.

Chaque séparation s'est faite sur les 7 même phrases que précédemment pour chaque combinaison de R locuteurs de référence pour chaque locuteur cible.

On obtient alors un tableau duquel on distingue :
— la moyenne des résultats pour les cas où le locuteur cible est parmi les locuteur de référence (noté ci-dessous Cible vs. Cible)
— la moyenne des résultats pour lesquels le locuteur cible n'est pas parmi les locuteurs de référence qui sont du même sexe que le locuteur cible (Cible vs. autres Références même sexe)
— la moyenne des résultats pour lesquels le locuteur cible n'est pas parmi les locuteurs de référence sans hypothèse sur leur genre (Cible vs. autres Références)

En raison de la combinatoire élevée de la contrainte d'identité, nous ne l'avons testé que pour un SNR de 0 dB.

3.2 Scores obtenus

3.2.1 Rappel des résultats pour l'algorithme sans déformation

On rappelle dans un premier temps les résultats obtenus sur ce corpus par l'algorithme existant (sans déformation) pour différents niveau de SNR : -6 dB, 0 dB et + 6 dB. Ces résultats seront désignés par Ref. par la suite.

	SDR			PESQ		
SNR	-6dB	0dB	+6dB	-6dB	0dB	+6dB
Cible vs. Cible	5.6	9.8	12.5	2.0	2.3	2.6
Cible F vs. autre Référence F	5.5	9.5	11.8	2.0	2.3	2.5
Cible H vs. autre Référence H	4.7	8.5	10.8	1.9	2.2	2.4
Cible vs. autre Référence	4.8	8.7	10.8	1.9	2.2	2.4

TABLE 3 – Résultats de référence pour une séparation à partir du modèle source/filtre - cas où $\mathbf{D}$ est fixée à $\mathbb{1}$

Comme on pouvait s'y attendre, les résultats sont meilleurs lorsque le locuteur sur lequel a lieu la séparation (la cible) est le même que celui pour lequel a eu lieu l'apprentissage (la référence) que lorsqu'ils diffèrent. De manière plus surprenante et inexpliquée, les résultats sont meilleurs lorsque cible et référence sont deux femmes que lorsque ce sont deux hommes.

3.2.2 Résultats pour l'algorithme avec déformation

On donne ensuite les résultats obtenus pour l'algorithme avec déformation pour une seule référence (R=1) sous différentes contraintes. On initialise dans tous les cas **D** aléatoirement, et on utilise dans tous les cas une contrainte de lissage spectral avec un poids $\lambda_{lissage} = 10$, qui est le poids pour lequel nous avons obtenu les meilleurs résultats.

On notera chaque contrainte de la façon suivante :
— Def. I pour la contrainte de petit gain
— Def. II pour la contrainte de corrélation maximale [1]
— Def. III pour la contrainte de réciprocité
La contrainte de décorrélation entre les bases du filtre et les bases de bruit n'a pas donné de résultats intéressants, ils n'apparaissent donc pas ici.

On obtient alors pour les différents SNR :

SNR = -6 dB

	SDR				PESQ			
	Ref	Def I	Def II	Def III	Ref	Def I	Def II	Def III
Cible vs. Cible	**5.6**	2.7	-	4.1	2.0	1.8	-	1.9
Cible F vs. autre Référence F	**5.5**	2.5	-	4.2	2.0	1.7	-	1.9
Cible H vs. autre Référence H	**4.7**	2.7	-	3.5	1.9	1.8	-	1.8
Cible vs. autre Référence	**4.8**	2.6	-	3.8	1.9	1.8	-	1.9

TABLE 4 – Résultats SNR= -6dB pour une séparation à partir du modèle source/filtre avec déformation contrainte. (Def I : contrainte de petite déformation - Def II : contrainte de corrélation maximale - Def III : contrainte de réciprocité)

SNR = 0 dB

	SDR				PESQ			
	Ref	Def I	Def II	Def III	Ref	Def I	Def II	Def III
Cible vs. Cible	**9.8**	8.1	8.1	9.7	2.3	2.2	2.2	**2.4**
Cible F vs. autre Référence F	9.5	8.4	8.4	**9.9**	2.3	2.2	2.2	**2.4**
Cible H vs. autre Référence H	8.5	7.5	7.2	**9.0**	2.3	2.2	2.2	**2.4**
Cible vs. autre Référence	8.7	7.9	7.9	**9.3**	2.3	2.2	2.2	**2.4**

TABLE 5 – Résultats SNR= 0dB pour une séparation à partir du modèle source/filtre avec déformation contrainte. (Def I : contrainte de petite déformation - Def II : contrainte de corrélation maximale - Def III : contrainte de réciprocité)

1. Le calcul de la contrainte de corrélation maximale est très long en raison de la double multiplication de matrices $F \times F$ (1025×1025 ici, soit 10h pour les 700 phrases par locuteur cible). Faute de temps, les calculs ne portent que sur deux locuteurs cibles (un homme et une femme) pour le SNR = 0 dB

Nous avons également testé pour le SNR = 0 dB la contrainte de réciprocité sans lissage ($\lambda_{lissage}$ = 0) en initialisant tous les termes de **D** à 1. On a obtenu des résultats légèrement moins bons qu'avec une initialisation aléatoire et lissage avec $\lambda_{lissage}$ = 10 (par exemple, on a obtenu un SDR Cible vs. Cible de 9.4 et Cible vs. autre Référence de 9.1).

SNR = +6 dB

	SDR				PESQ			
	Ref	Def I	Def II	Def III	Ref	Def I	Def II	Def III
Cible vs. Cible	12.5	13.2	-	**13.6**	2.6	2.6	-	**2.7**
Cible F vs. autre Référence F	11.8	13.5	-	**13.9**	2.5	2.5	-	**2.5**
Cible H vs. autre Référence H	10.8	12.6	-	**12.8**	2.4	2.6	-	**2.7**
Cible vs. autre Référence	10.8	12.9	-	**13.1**	2.4	2.5	-	**2.6**

TABLE 6 – Résultats SNR= +6dB pour une séparation à partir du modèle source/filtre avec déformation contrainte. (Def I : contrainte de petite déformation - Def II : contrainte de corrélation maximale - Def III : contrainte de réciprocité)

3.2.3 Resultats avec la déformation et de multiples références

On s'intéresse à présent aux résultats des séparations avec un nombre de références > 1, avec la contrainte d'identité. Les meilleurs résultats sont obtenus pour un poids sur cette contrainte de λ_{id} = 10.

Comme indiqué précédemment, en raison de la combinatoire importante à tester, seuls les résultats pour un SNR de 0dB ont été calculés.

	Ref	SDR					Ref	PESQ				
R	1	1	2	3	4	5	1	1	2	3	4	5
Cible vs. Cible	**9.8**	9.7	9.7	9.4	9.4	9.1	2.3	**2.4**	2.3	2.3	2.3	2.3
Cible F vs. autre Référence F	9.5	**9.9**	8.9	-	-	-	2.3	**2.4**	2.3	-	-	-
Cible H vs. autre Référence H	8.5	**9.0**	8.9	-	-	-	2.3	**2.4**	2.2	-	-	-
Cible vs. autre Référence	8.7	**9.3**	9.1	8.9	8.9	8.7	2.3	**2.4**	2.2	2.3	2.3	2.3

TABLE 7 – Résultats pour SNR=0 dB, avec R=2 à 5

Pour R>2, il n'y a pas de combinatoire où toutes les références sont uniquement des hommes ou uniquement des femmes. Les items "Cible F. vs autres Références F." et "Cible M. vs autres Références M." ne sont donc pas applicables.

3.3 Discussion des résultats

3.3.1 Contraintes de petite déformation

Pour un SNR de 0 dB, les contraintes de petit gain et de maximum de corrélation entre bases de même phonème donnent des résultats meilleurs que lorsque la matrice **D** de déformation est laissée libre sans contrainte (cf. table 2 et table 5). Néanmoins, elles ne permettent pas d'atteindre les scores de référence obtenus pour $\mathbf{W}^{\Phi}$ fixée sans déformation. La contrainte de petit gain donne également de moins bons résultats

que le cas sans déformation à -6dB, mais meilleurs à + 6 dB. Les hypothèses sous-jacentes à la contrainte de petit gain correspondent donc bien à une réalité physique, même si l'effet s'atténue lorsque le bruit augmente.

3.3.2 Contrainte de réciprocité

Pour un SNR de -6 dB, le modèle de déformation sous contrainte de réciprocité donne des résultats moins bons que les résultats de référence sans déformation, mais meilleurs que les contraintes de petit gain (cf. table 4).

Ce modèle donne des résultats meilleurs que toutes les autres méthodes pour un SNR de 0 dB (cf. table 5) et pour un SNR de +6dB (cf. table 6), que le locuteur de référence soit ou non le locuteur cible.

On observe de manière logique que les résultats sont meilleurs lorsque le locuteur cible est le locuteur de référence pour lequel a eu lieu l'apprentissage. Mais on observe également que l'écart entre les résultats "Cible vs. Cible" et "Cible vs. autre Référence" est moins important que pour l'algorithme sans déformation : pour un SDR de 0 dB, on observe un écart de SDR de 1.1 pour l'algorithme sans déformation vs. un écart de 0.4 pour l'algorithme avec déformation.

En d'autres termes, l'algorithme avec déformation donne de meilleurs résultats que celui sans adaptation pour les cas où le locuteur cible est inconnu : l'adaptation remplit bien le rôle que l'on en attendait, à savoir une capacité à effectuer une séparation pour des locuteurs inconnus à un niveau proche de ceux observé pour des locuteurs connus [2].

On observe également que, lorsque le locuteur cible n'est *pas* le locuteur de référence mais est de même sexe, on obtient de meilleurs résultats si c'est une femme que si c'est un homme. Nous n'avons pas d'interprétation à proposer pour l'expliquer.

On peut enfin faire remarquer que les temps de calculs sont largement améliorés par ce nouvel algorithme : en raison de calculs de contraintes plus simples et l'utilisation de la multiplication terme à terme, l'algorithme avec contrainte de réciprocité converge 2.6 fois plus rapidement que l'algorithme initial (30 mn vs. 1h16 mn) pour le traitement des 700 phrases utilisées pour un locuteur cible.

3.3.3 Contrainte d'identité

Alors que l'introduction d'une contrainte de cohérence sous forme de contrainte de réciprocité a un effet positif sur la convergence de la NMF, l'introduction d'équations supplémentaires ne renforce pas cet effet. Au vu des résultats, il apparait que c'est même l'inverse, dans la mesure où plus il y a de références, moins les résultats sont bons.

Néanmoins, dans la mesure où la contrainte d'identité n'a pas pu être testée sur l'ensemble des locuteurs, et que des variations importantes sont observées d'un locuteur à l'autre en termes de résultats, il serait nécessaire de mener les calculs en entier pour pouvoir conclure définitivement.

Pour l'instant, il semble cependant que la contrainte d'identité n'apporte pas d'amélioration. Elle semble donc d'autant moins utile qu'elle complexifie les calculs en raison de la nécessité de procéder à un alignement des bases entre locuteurs en cas de phonèmes manquants d'une part, et est plus lente que la simple contrainte de réciprocité en raison du grand nombre de références à traiter d'autre part.

2. Des exemples audio sont disponibles sur http ://recherche.ircam.fr/anasyn/speechSeparation/examples.html

Conclusion et perspectives

Nous avons vu que la NMF semi-supervisée permet d'obtenir de bons résultats en séparation de la parole dès lors que le locuteur cible est celui pour lequel l'apprentissage a eu lieu au préalable. Les résultats sont moins bons lorsque l'apprentissage a eu lieu pour un locuteur différent. Il est donc apparu nécessaire d'adapter les enveloppes spectrales apprises sur un locuteur donné pour réaliser une séparation satisfaisante sur un autre locuteur inconnu.

En nous inspirant des différentes méthodes disponibles dans la littérature pour réaliser ce type d'adaptation, nous avons proposé un nouveau modèle permettant de décrire la déformation des enveloppes spectrales d'un locuteur à un autre. Nous avons également proposé deux types de contraintes - contraintes de petite déformation et contraintes de déformation cohérente - afin de pouvoir guider et contrôler l'adaptation au cours de la mise à jour par NMF et d'obtenir une séparation satisfaisante.

Nous avons alors vu que l'une des contraintes de déformation cohérente - que nous avons appelé contrainte de réciprocité - permet d'obtenir des résultats meilleurs que dans le cas sans déformation pour des SNR de 0 dB et de 6 dB, en particulier dès lors que le locuteur cible n'est pas celui pour lequel a eu lieu l'apprentissage. Il s'avère par ailleurs que cet algorithme est plus 2.5 fois plus rapide que l'algorithme sans déformation.

A ce stade de notre réflexion, il nous semble qu'il sera difficile d'améliorer ces résultats dans le cadre de la NMF en raison de son formalisme même qui nécessite une expression *analytique* des contraintes à imposer à la déformation. Or, en raison de la variabilité des enveloppes spectrales pour un même phonème d'un locuteur à l'autre, il parait illusoire d'espérer définir une contrainte garantissant la cohérence de toutes les adaptations pour tous les locuteurs d'un corpus de test. Nous avons donc proposé pour un éventuel travail à venir d'essayer de réaliser un apprentissage non seulement sur les enveloppes spectrales des locuteurs, mais sur les déformations observées d'un locuteur à l'autre.

A cet effet, deux pistes ont été envisagées pour réaliser un apprentissage de la cohérence d'une adaptation entre phonèmes : d'une part, une reformulation probabiliste du problème de séparation (e.g. par PLCA [Smaragdis et al., 2007]) qui permettrait d'obtenir l'adaptation la plus probable à partir d'une loi de distribution des déformations observées sur de multiples locuteurs ; d'autre part une reformulation du problème dans un formalisme proche de celui des réseaux de neurones (e.g. par deep-NMF [Le Roux et al., 2015]) qui permettrait de pouvoir bénéficier à la fois des capacités de modélisation de la NMF et des capacités d'apprentissage propres aux réseaux de neurones.

Annexe A Calculs des gradients des modèles de déformation

Rappel : on cherche à minimiser la β-divergence entre deux matrices $\mathbf{X}$ et $\mathbf{V}$ qui s'écrit :

$$d_\beta(\mathbf{X}|\mathbf{V}) = \sum_{i,j} \frac{1}{\beta(\beta-1)} (\mathbf{X}_{ij}^\beta + (\beta-1)\mathbf{V}_{ij}^\beta - \beta \mathbf{X}_{ij}\mathbf{V}_{ij}^{\beta-1})$$

On obtient la valeur de l'un des termes $\boldsymbol{\theta}$ de $\mathbf{V}$ minimisant cette β-divergence par une descente de gradient par rapport à $\boldsymbol{\theta}$. La mise à jour itérative à l'itération n de $\boldsymbol{\theta}$ s'écrit classiquement :

$$\boldsymbol{\theta}^{(n+1)} \leftarrow \boldsymbol{\theta}^{(n)} - \eta \nabla_{\boldsymbol{\theta}} d_\beta(\mathbf{X}|\mathbf{V})$$

où η est le pas de la descente de gradient. En notant $\nabla^+_{\boldsymbol{\theta}}$ et $\nabla^-_{\boldsymbol{\theta}}$ la somme des termes respectivement positifs et négatifs de $\nabla_{\boldsymbol{\theta}} d_\beta(\mathbf{X}|\mathbf{V})$, et sous l'hypothèse que tous les coefficients de la matrice $\boldsymbol{\theta}$ sont positifs, on peut choisir η tel que

$$\eta = \frac{\boldsymbol{\theta}^{(n)}}{\nabla^+_{\boldsymbol{\theta}}}$$

La formule de mise à jour d'une itération à l'autre s'écrit alors

$$\boldsymbol{\theta}^{(n+1)} \leftarrow \boldsymbol{\theta}^{(n)} \otimes \frac{\nabla^-_{\boldsymbol{\theta}}}{\nabla^+_{\boldsymbol{\theta}}}$$

On utilisera ce principe y compris lorsque la fonction de coût à minimiser sera la β-divergence et d'autres termes de contraintes en séparant les termes positifs et négatifs du gradient de la fonction de coût total, bien qu'il n'y ait plus de preuve de convergence dans ces cas là.

Par ailleurs, la dérivée de cette expression par rapport à l'un des termes constituant $\mathbf{V}$ s'écrit donc :

$$\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \theta_{pq}} = \sum_{i,j} \mathbf{V}_{ij}^{\beta-1} \frac{\partial \mathbf{V}_{ij}}{\partial \theta_{pq}} - \sum_{i,j} \mathbf{X}_{ij}\mathbf{V}_{ij}^{\beta-2} \frac{\partial \mathbf{V}_{ij}}{\partial \theta_{pq}} \tag{30}$$

A.1 $\mathbf{V} = \mathbf{WH}$

$$\frac{\partial \mathbf{V}_{ij}}{\partial \mathbf{W}_{pq}} = \frac{\partial (\mathbf{WH})_{ij}}{\partial \mathbf{W}_{pq}} = \sum_k \frac{\partial \mathbf{W}_{ik}}{\partial \mathbf{W}_{pq}} \mathbf{H}_{kj} = \sum_k \delta_{ip}\delta_{kq}\mathbf{H}_{kj} = \delta_{ip}\mathbf{H}_{qj}$$

En injectant dans (30) :

$$\begin{aligned}
\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \mathbf{W}_{pq}} &= \sum_{i,j} \mathbf{V}_{ij}^{\beta-1}\delta_{ip}\mathbf{H}_{qj} - \sum_{i,j} \mathbf{X}_{ij}\mathbf{V}_{ij}^{\beta-2}\delta_{ip}\mathbf{H}_{qj} \\
&= \sum_j \mathbf{V}_{pj}^{\beta-1}\mathbf{H}_{qj} - \sum_j \mathbf{X}_{pj}\mathbf{V}_{pj}^{\beta-2}\mathbf{H}_{qj} \\
&= \sum_j \mathbf{V}_{pj}^{\beta-1}\mathbf{H}^T_{jq} - \sum_j \mathbf{X}_{pj}\mathbf{V}_{pj}^{\beta-2}\mathbf{H}^T_{jq} \\
&= (\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T)_{pq} - ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T)_{pq}
\end{aligned}$$

$$\frac{\partial \mathbf{V}_{ij}}{\partial \mathbf{H}_{pq}} = \frac{\partial (\mathbf{WH})_{ij}}{\partial \mathbf{H}_{pq}} = \sum_k \mathbf{W}_{ik} \frac{\partial \mathbf{H}_{kj}}{\partial \mathbf{H}_{pq}} = \sum_k \mathbf{W}_{ik} \delta_{kp} \delta_{jq} = \delta_{jq} \mathbf{W}_{ip}$$

En injectant dans (30) :

$$\begin{aligned}
\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \mathbf{H}_{pq}} &= \sum_{i,j} \mathbf{V}_{ij}^{\beta-1} \delta_{jq} \mathbf{W}_{ip} - \sum_{i,j} \mathbf{X}_{ij} \mathbf{V}_{ij}^{\beta-2} \delta_{jq} \mathbf{W}_{ip} \\
&= \sum_i \mathbf{V}_{iq}^{\beta-1} \mathbf{W}_{ip} - \sum_i \mathbf{X}_{iq} \mathbf{V}_{iq}^{\beta-2} \mathbf{W}_{ip} \\
&= \sum_i \mathbf{W}_{pi}^T \mathbf{V}_{iq}^{\beta-1} - \sum_i \mathbf{W}_{pi}^T \mathbf{X}_{iq} \mathbf{V}_{iq}^{\beta-2} \\
&= (\mathbf{W}^T \mathbf{V}^{\cdot(\beta-1)})_{pq} - (\mathbf{W}^T (\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)}))_{pq}
\end{aligned}$$

Les règles de mise à jour sont donc :

$$\mathbf{W}^{(i+1)} = \mathbf{W}^{(i)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)}) \mathbf{H}^T}{\mathbf{V}^{\cdot(\beta-1)} \mathbf{H}^T}$$

$$\mathbf{H}^{(i+1)} = \mathbf{H}^{(i)} \otimes \frac{\mathbf{W}^T (\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})}{\mathbf{W}^T \mathbf{V}^{\cdot(\beta-1)}}$$

A.2 $\mathbf{V} = (\mathbf{W} + \mathbf{G})\mathbf{H}$

On est dans le cas où **G** joue le rôle de **W** dans le cas classique :

$$\frac{\partial \mathbf{V}_{ij}}{\partial \mathbf{G}_{pq}} = \frac{\partial [(\mathbf{W} + \mathbf{G})\mathbf{H}]_{ij}}{\partial \mathbf{G}_{pq}} = \sum_k \frac{\partial \mathbf{G}_{ik}}{\partial \mathbf{G}_{pq}} \mathbf{H}_{kj} = \sum_k \delta_{ip} \delta_{kq} \mathbf{H}_{kj} = \delta_{ip} \mathbf{H}_{qj} \tag{31}$$

En injectant dans (30) :

$$\begin{aligned}
\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \mathbf{G}_{pq}} &= \sum_{i,j} \mathbf{V}_{ij}^{\beta-1} \delta_{ip} \mathbf{H}_{qj} - \sum_{i,j} \mathbf{X}_{ij} \mathbf{V}_{ij}^{\beta-2} \delta_{ip} \mathbf{H}_{qj} \\
&= \sum_j \mathbf{V}_{pj}^{\beta-1} \mathbf{H}_{qj} - \sum_j \mathbf{X}_{pj} \mathbf{V}_{pj}^{\beta-2} \mathbf{H}_{qj} \\
&= \sum_j \mathbf{V}_{pj}^{\beta-1} \mathbf{H}_{jq}^T - \sum_j \mathbf{X}_{pj} \mathbf{V}_{pj}^{\beta-2} \mathbf{H}_{jq}^T \\
&= [\mathbf{V}^{\cdot(\beta-1)} \mathbf{H}^T]_{pq} - [(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)}) \mathbf{H}^T]_{pq} \\
&= (\mathbf{V}^{\cdot(\beta-1)} \mathbf{H}^T)_{pq} - ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)}) \mathbf{H}^T)_{pq}
\end{aligned}$$

Seulement si **G** est non-négative, les règles de mise à jour s'écrivent donc :

$$\mathbf{G}^{(i+1)} = \mathbf{G}^{(i)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)}) \mathbf{H}^T}{\mathbf{V}^{\cdot(\beta-1)} \mathbf{H}^T}$$

A.3 $\mathbf{V} = (\mathbf{D} \otimes \mathbf{W})\mathbf{H}$

$$\frac{\partial \mathbf{V}_{ij}}{\partial \mathbf{D}_{pq}} = \frac{\partial [(\mathbf{D} \otimes \mathbf{W})\mathbf{H}]_{ij}}{\partial \mathbf{D}_{pq}} = \sum_k \frac{\partial \mathbf{D}_{ik}}{\partial \mathbf{D}_{pq}} \mathbf{W}_{ik}\mathbf{H}_{kj} = \sum_k \delta_{ip}\delta_{kq}\mathbf{W}_{ik}\mathbf{H}_{kj} = \delta_{ip}\mathbf{W}_{iq}\mathbf{H}_{qj} \tag{32}$$

En injectant dans (30), il vient :

$$\begin{aligned}
\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \mathbf{D}_{pq}} &= \sum_{i,j} \mathbf{V}_{ij}^{\beta-1}\delta_{ip}\mathbf{W}_{iq}\mathbf{H}_{qj} - \sum_{i,j} \mathbf{X}_{ij}\mathbf{V}_{ij}^{\beta-2}\delta_{ip}\mathbf{W}_{iq}\mathbf{H}_{qj} \\
&= \sum_j \mathbf{V}_{pj}^{\beta-1}\mathbf{W}_{pq}\mathbf{H}_{qj} - \sum_j \mathbf{X}_{pj}\mathbf{V}_{pj}^{\beta-2}\mathbf{W}_{pq}\mathbf{H}_{qj} \\
&= \mathbf{W}_{pq}\sum_j \mathbf{V}_{pj}^{\beta-1}\mathbf{H}_{jq}^T - \mathbf{W}_{pq}\sum_j \mathbf{X}_{pj}\mathbf{V}_{pj}^{\beta-2}\mathbf{H}_{jq}^T \\
&= \mathbf{W}_{pq}[\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T]_{pq} - \mathbf{W}_{pq}[(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T]_{pq} \\
&= [\mathbf{W} \otimes (\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T)]_{pq} - [\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T)]_{pq}
\end{aligned}$$

La dérivation par rapport à $\mathbf{H}$ s'écrit simplement par analogie avec le cas classique :

$$\frac{\partial d_\beta(\mathbf{X}|\mathbf{V})}{\partial \mathbf{H}_{pq}} = \frac{\partial d_\beta(\mathbf{X}|(\mathbf{D} \otimes \mathbf{W})\mathbf{H})}{\partial \mathbf{H}_{pq}} = [(\mathbf{D} \otimes \mathbf{W})^T\mathbf{V}^{\cdot(\beta-1)}]_{pq} - [(\mathbf{D} \otimes \mathbf{W})^T(\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})]_{pq}$$

Si $\mathbf{D}$ est non-négative, les règles de mise à jour s'écrivent donc :

$$\mathbf{D}^{(i+1)} = \mathbf{D}^{(i)} \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes ((\mathbf{D} \otimes \mathbf{W})\mathbf{H})^{\cdot(\beta-2)})\mathbf{H}^T)}{\mathbf{W} \otimes (((\mathbf{D} \otimes \mathbf{W})\mathbf{H})^{\cdot(\beta-1)}\mathbf{H}^T)}$$

$$\mathbf{H}^{(i+1)} = \mathbf{H}^{(i)} \otimes \frac{(\mathbf{D} \otimes \mathbf{W})^T(\mathbf{X} \otimes ((\mathbf{D} \otimes \mathbf{W})\mathbf{H})^{\cdot(\beta-2)})}{(\mathbf{D} \otimes \mathbf{W})^T((\mathbf{D} \otimes \mathbf{W})\mathbf{H})^{\cdot(\beta-1)}}$$

Annexe B Calculs des gradients des contraintes utilisées

B.1 Contraintes de petite déformation

B.1.1 Contrainte de petit gain

On cherche à minimiser le coût :

$$\mathscr{C}(\mathbf{D}) = \|\mathbf{D}\otimes\mathbf{W}-\mathbf{W}\|^2 = \sum_{i,j}|\mathbf{D}\otimes\mathbf{W}-\mathbf{W}|^2_{ij}$$

On a :

$$\begin{aligned}\frac{\partial\mathscr{C}(\mathbf{D})}{\partial\mathbf{D}_{pq}} &= 2\sum_{i,j}(\mathbf{D}\otimes\mathbf{W}-\mathbf{W})_{ij}\frac{\partial\mathbf{D}_{ij}\mathbf{W}_{ij}}{\partial\mathbf{D}_{pq}}\\ &= 2\sum_{i,j}\mathbf{W}^2_{ij}(\mathbf{D}_{ij}-1)\delta_{ip}\delta_{jq}\\ &= 2\mathbf{W}^2_{pq}(\mathbf{D}_{pq}-1)\end{aligned}$$

ce qui s'écrit sous forme matricelle :

$$\frac{\partial\mathscr{C}(\mathbf{D})}{\partial\mathbf{D}} = 2\mathbf{D}\otimes\mathbf{W}\otimes\mathbf{W}-2\mathbf{W}\otimes\mathbf{W}$$

B.1.2 Contrainte de maximum de corrélation entre mêmes phonèmes

On cherche à maximiser (donc minimiser son opposé) le coût :

$$\mathscr{C}(\mathbf{D}) = \|(\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}\|^2 = \sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})^2_{ij}$$

On a :

$$\begin{aligned}\frac{\partial\mathscr{C}(\mathbf{D})}{\partial\mathbf{D}_{pq}} &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}\frac{\partial((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}}{\partial\mathbf{D}_{pq}}\\ &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}\sum_k\frac{\partial\mathbf{D}_{ki}}{\partial\mathbf{D}_{pq}}\mathbf{W}_{ki}\mathbf{W}_{kj}\\ &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}\sum_k\frac{\partial\mathbf{D}_{ki}}{\partial\mathbf{D}_{pq}}\mathbf{W}_{ki}\mathbf{W}_{kj}\\ &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}\sum_k\delta_{kp}\delta_{iq}\mathbf{W}_{ki}\mathbf{W}_{kj}\\ &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{ij}\delta_{iq}\mathbf{W}_{pq}\mathbf{W}_{pj}\\ &= 2\sum_j((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W})_{qj}\mathbf{W}_{pq}\mathbf{W}_{pj}\\ &= 2\mathbf{W}_{pq}(\mathbf{W}\mathbf{W}^T(\mathbf{D}\otimes\mathbf{W}))_{pq}\\ &= 2(\mathbf{W}\otimes(\mathbf{W}\mathbf{W}^T(\mathbf{D}\otimes\mathbf{W})))_{pq}\end{aligned}$$

ce qui s'écrit sous forme matricelle :

$$\frac{\partial \mathscr{C}(\mathbf{D})}{\partial \mathbf{D}} = 2\mathbf{W} \otimes (\mathbf{W}\mathbf{W}^T(\mathbf{D} \otimes \mathbf{W}))$$

B.1.3 Contrainte de lissage spectral

On observe que la reconstruction de $\mathbf{D} \otimes \mathbf{W}$ est très hachée si l'on initialise $\mathbf{D}$ aléatoirement. On introduit donc une contrainte de lissage. On prend comme coût $\mathscr{C}(\mathbf{D}) = \sum_{ij} |(\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j}|^2$.

La dérivation par rapport à $\mathbf{D}$ s'écrit :

$$\begin{aligned}
\frac{\partial \mathscr{C}(\mathbf{D})}{\partial \mathbf{D}} &= 2\sum_{ij}((\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j}) \frac{\partial (\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j}}{\partial \mathbf{D}_{pq}} \\
&= 2\sum_{ij}((\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j})[\frac{\partial \mathbf{D}_{i+1,j}}{\partial \mathbf{D}_{pq}}\mathbf{W}_{i+1,j} - \frac{\partial \mathbf{D}_{i,j}}{\partial \mathbf{D}_{pq}}\mathbf{W}_{i,j}] \\
&= 2\sum_{ij}((\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j})\mathbf{W}_{i+1,j}\delta_{i+1,p}\delta_{j,q} - 2\sum_{ij}((\mathbf{D} \otimes \mathbf{W})_{i+1,j} - (\mathbf{D} \otimes \mathbf{W})_{i,j})\mathbf{W}_{i,j}\delta_{i,p}\delta_{j,q} \\
&= 2((\mathbf{D} \otimes \mathbf{W})_{pq} - (\mathbf{D} \otimes \mathbf{W})_{p-1,q})\mathbf{W}_{pq} - 2((\mathbf{D} \otimes \mathbf{W})_{p+1,q} - (\mathbf{D} \otimes \mathbf{W})_{pq})\mathbf{W}_{pq} \\
&= 4((\mathbf{D} \otimes \mathbf{W})_{pq}\mathbf{W}_{pq} - 2((\mathbf{D} \otimes \mathbf{W})_{p-1,q} + (\mathbf{D} \otimes \mathbf{W})_{p+1,q})\mathbf{W}_{pq}
\end{aligned}$$

Ce qui s'écrit sous forme matricielle :

$$\frac{\partial \mathscr{C}(\boldsymbol{D})}{\partial \boldsymbol{D}} = 4\mathbf{D} \otimes \mathbf{W} \otimes \mathbf{W} - 2((\mathbf{D} \otimes \mathbf{W})^{\downarrow} + (\mathbf{D} \otimes \mathbf{W})^{\uparrow}) \otimes \mathbf{W}$$

où pour $\forall p, q$, $(\mathbf{D} \otimes \mathbf{W})^{\downarrow}_{pq} = (\mathbf{D} \otimes \mathbf{W})_{p-1,q}$ et $(\mathbf{D} \otimes \mathbf{W})^{\uparrow}_{pq} = (\mathbf{D} \otimes \mathbf{W})_{p+1,q}$.

B.2 Contraintes de déformation cohérente

B.2.1 Contraintes de décorrélation avec le bruit

On reprend [Kitamura et al., 2013] pour introduire une contrainte de décorrélation entre les bases déformées de $\mathbf{D} \otimes \mathbf{W}^{\Phi}$ et $\mathbf{W}_N$ dont l'objectif est de séparer le signal qui relève du bruit du signal utile. Dans ce qui suit, on ne marque pas l'exposant Φ pour alléger l'écriture. Le coût de la contrainte s'écrit :

$$\mathscr{C}(\mathbf{D}, \mathbf{W}_N) = \|(\mathbf{D} \otimes \mathbf{W})^T \mathbf{W}_N\|_F^2 = \sum_{i,j} |(\mathbf{D} \otimes \mathbf{W})^T \mathbf{W}_N|_{ij}^2$$

La dérivation par rapport à $\mathbf{D}$ s'écrit :

$$\begin{aligned}
\frac{\partial \mathscr{C}(\mathbf{D},\mathbf{W}_N)}{\partial \mathbf{D}_{pq}} &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{ij}\sum_k \frac{\partial((\mathbf{D}\otimes\mathbf{W})_{ki}\mathbf{W}_{N,kj})}{\partial \mathbf{D}_{pq}} \\
&= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{ij}\sum_k \delta_{kp}\delta_{iq}\mathbf{W}_{ki}\mathbf{W}_{N,kj} \\
&= 2\sum_{j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{qj}\mathbf{W}_{pq}\mathbf{W}_{N,pj} \\
&= 2\mathbf{W}_{pq}(\mathbf{W}_N\mathbf{W}_N^T(\mathbf{D}\otimes\mathbf{W}))_{pq} \\
&= 2[\mathbf{W}\otimes(\mathbf{W}_N\mathbf{W}_N^T(\mathbf{D}\otimes\mathbf{W}))]_{pq}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathscr{C}(\mathbf{D},\mathbf{W}_N)}{\partial \mathbf{W}_{N,pq}} &= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{ij}\sum_k \frac{\partial((\mathbf{D}\otimes\mathbf{W})_{ki}\mathbf{W}_{N,kj})}{\partial \mathbf{W}_{N,pq}} \\
&= 2\sum_{i,j}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{ij}\sum_k (\mathbf{D}\otimes\mathbf{W})_{ki}\delta_{kp}\delta_{jq} \\
&= 2\sum_{i}((\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)_{iq}(\mathbf{D}\otimes\mathbf{W})_{pi} \\
&= 2[(\mathbf{D}\otimes\mathbf{W})(\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)]_{pq}
\end{aligned}$$

Ce qui s'écrit sous forme matricielle :

$$\frac{\partial \mathscr{C}(\mathbf{D},\mathbf{W}_N)}{\partial \mathbf{D}} = 2\mathbf{W}\otimes(\mathbf{W}_N\mathbf{W}_N^T(\mathbf{D}\otimes\mathbf{W})) \qquad \frac{\partial \mathscr{C}(\mathbf{D},\mathbf{W}_N)}{\partial \mathbf{W}_N} = 2(\mathbf{D}\otimes\mathbf{W})(\mathbf{D}\otimes\mathbf{W})^T\mathbf{W}_N)$$

B.2.2 Contrainte de réciprocité

On introduit une déformation réciproque entre les bases apprises connues et les bases déformées à optimiser en considérant le système suivant :

$$\begin{cases}
\mathbf{V}_{0/r} &= \mathbf{W}^{ex}\mathbf{H}_0^{ex}\otimes(\mathbf{D}_{0/r}\otimes\mathbf{W}_r^{\Phi})\mathbf{H}_0^{\Phi} + \mathbf{W}^N\mathbf{H}^N \\
\mathbf{V}_r &= \mathbf{W}^{ex}\mathbf{H}_r^{ex}\otimes(\mathbf{D}_r\otimes\mathbf{D}_{0/r}\otimes\mathbf{W}_r^{\Phi})\mathbf{H}_r^{\Phi}
\end{cases}$$

On impose la contrainte de minimiser :

$$\mathscr{C} = \sum_{r=1}^{R} d_\beta(\mathbf{V}_{0/r}|\mathbf{X}_0) + d_\beta(\mathbf{V}_r|\mathbf{X}_r)$$

En reprenant le gradient du cas général de déformation de l'annexe A, il vient :

$$\begin{aligned}
\frac{\partial \mathscr{C}(\mathbf{D}_{0/r},\mathbf{D}_r)}{\partial \mathbf{D}_{0/r}} &= \mathbf{W}_r^{\Phi}\otimes((\mathbf{V}_0^{ex}\otimes\mathbf{V}_{0/r}^{(\beta-1)})\mathbf{H}_0^T) + (\mathbf{D}_r\otimes\mathbf{W}_r^{\Phi})\otimes((\mathbf{V}_r^{ex}\otimes\mathbf{V}_r^{(\beta-1)})\mathbf{H}_r^T) - \\
&\qquad \mathbf{W}_r^{\Phi}\otimes((\mathbf{V}_0^{ex}\otimes\mathbf{X}_0\otimes\mathbf{V}_{0/r}^{(\beta-2)})\mathbf{H}_0^T) + (\mathbf{D}_r\otimes\mathbf{W}_r^{\Phi})\otimes((\mathbf{V}_0^{ex}\otimes\mathbf{X}_r\otimes\mathbf{V}_r^{(\beta-2)})\mathbf{H}_r^T)
\end{aligned}$$

et

$$\frac{\partial \mathscr{C}(\mathbf{D}_{0/r},\mathbf{D}_r)}{\partial \mathbf{D}_r} = (\mathbf{D}_{0/r}\otimes\mathbf{W}_r^{\Phi})\otimes((\mathbf{V}_r^{ex}\otimes\mathbf{V}_r^{(\beta-1)})\mathbf{H}_r^T) - (\mathbf{D}_{0/r}\otimes\mathbf{W}_r^{\Phi})\otimes((\mathbf{V}_r^{ex}\otimes\mathbf{X}_r\otimes\mathbf{V}_r^{(\beta-2)})\mathbf{H}_r^T)$$

B.2.3 Contraintes d'identité

On rappelle le système à plusieurs références :

$$\begin{cases} \mathbf{V}_{0/1} &= (\mathbf{D}_{0/1} \otimes \mathbf{W}_1)\mathbf{H}_0 + \mathbf{W}^N\mathbf{H}^N \\ & \dots \\ \mathbf{V}_{0/r} &= (\mathbf{D}_{0/r} \otimes \mathbf{W}_r)\mathbf{H}_0 + \mathbf{W}^N\mathbf{H}^N \\ & \dots \\ \mathbf{V}_{0/R} &= (\mathbf{D}_{0/R} \otimes \mathbf{W}_R)\mathbf{H}_0 + \mathbf{W}^N\mathbf{H}^N \end{cases}$$

On cherche à minimiser :

$$\mathscr{C}_id(\mathbf{D}_{0/\rho}) = \sum_{\rho=r}^{R} \|\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho\|_F^2 \qquad \forall r \in \{1...R\}$$

Or :

$$\begin{aligned}
\frac{\partial \sum_{\rho=1}^{R} \mathscr{C}_id(\mathbf{D}_{0/\rho})}{\partial \mathbf{D}_{0/r,pq}} &= \sum_{\rho=1}^{R}\sum_{s=\rho}^{R} \frac{\partial \|\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s\|_F^2}{\partial \mathbf{D}_{0/r,pq}} = \sum_{\rho=1}^{R}\sum_{s=\rho}^{R}\sum_{i,j} \frac{\partial |\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s|_{ij}^2}{\partial \mathbf{D}_{0/r,pq}} \\
&= 2\sum_{\rho=1}^{R}\sum_{s=\rho}^{R}\sum_{i,j} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{ij} \frac{\partial (\mathbf{D}_{0/\rho,ij}\mathbf{W}_{\rho,ij} - \mathbf{D}_{0/s,ij}\mathbf{W}_{s,ij})}{\partial \mathbf{D}_{0/r,pq}} \\
&= 2\sum_{\rho=1}^{R}\sum_{s=\rho}^{R}\sum_{i,j} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{ij} (\delta_{\rho r}\delta_{ip}\delta_{jq}\mathbf{W}_{\rho,ij} - \delta_{sr}\delta_{ip}\delta_{jq}\mathbf{W}_{s,ij}) \\
&= 2\sum_{\rho=1}^{R}\sum_{s=\rho}^{R} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{pq} (\delta_{\rho r}\mathbf{W}_{\rho,pq} - \delta_{sr}\mathbf{W}_{s,pq}) \\
&= 2\sum_{\rho=1}^{R} \delta_{\rho r}\mathbf{W}_{\rho,pq} \sum_{s=\rho}^{R} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{pq} - 2\sum_{\rho=1}^{R}\sum_{s=\rho}^{R} \delta_{sr}\mathbf{W}_{s,pq} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{pq} \\
&= 2\mathbf{W}_{r,pq} \sum_{s=r}^{R} (\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/s} \otimes \mathbf{W}_s)_{pq} - 2\sum_{\rho=1}^{r} (\mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho - \mathbf{D}_{0/r} \otimes \mathbf{W}_r)_{pq}\mathbf{W}_{r,pq} \\
&= 2\mathbf{W}_{r,pq} \sum_{\rho=r}^{R} (\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho)_{pq} + 2\mathbf{W}_{r,pq} \sum_{\rho=1}^{r} (\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho)_{pq} \\
&= 2\mathbf{W}_{r,pq} \sum_{\rho=1}^{R} (\mathbf{D}_{0/r} \otimes \mathbf{W}_r - \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho)_{pq}
\end{aligned}$$

Soit sous forme matricielle :

$$\frac{\partial \sum_{\rho=1}^{R} \mathscr{C}_id(\mathbf{D}_{0/\rho})}{\partial \mathbf{D}_{0/r}} = 2R\mathbf{W}_r \otimes \mathbf{D}_{0/r} \otimes \mathbf{W}_r - 2\mathbf{W}_r \otimes \sum_{\rho=1}^{R} \mathbf{D}_{0/\rho} \otimes \mathbf{W}_\rho$$

Annexe C Equivalence de la NMF des déformations additive et multiplicative terme à terme

On considère le modèle de déformation additif :

$$\mathbf{X} \simeq \mathbf{V} = (\mathbf{W} + \mathbf{G})\mathbf{H}$$

Le coût total à minimiser s'écrit en fonction de $\mathbf{G}$:

$$\mathscr{C} = d_\beta(\mathbf{V}|\mathbf{X}) + \mathscr{C}_\mathbf{G} + \mathbf{cste}$$

On a alors pour $\forall\, p, q \in \{1...F\} \times \{1...K\}$ (le détail des calculs est donné en annexe A) :

$$\frac{\partial \mathscr{C}}{\partial \mathbf{G}_{pq}} = (\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T)_{pq} - ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T)_{pq} + \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{G}_{pq}}$$

Considérons à présent l'écriture équivalente de $\mathbf{V}$ avec une déformation multiplicative terme à terme :

$$\mathbf{V} = (\mathbf{W} + \mathbf{G})\mathbf{H} = (\mathbf{D} \otimes \mathbf{G})\mathbf{H} \qquad \text{avec } \mathbf{D} = \mathbb{1} + \mathbf{G}./\mathbf{W}$$

où $\mathbb{1}$ est la matrice dont tous les coefficients sont à 1 et où le symbole ./ représente une division terme à terme.

La fonction de coût à minimiser peut alors s'écrire en fonction de $\mathbf{D}$ seulement (en remplaçant $\mathbf{G}$ par son expression en fonction de $\mathbf{D}$) :

$$\mathscr{C} = d_\beta(\mathbf{V}|\mathbf{X}) + \mathscr{C}_\mathbf{D} + \mathbf{cste}$$

et on peut alors exprimer l'expression du gradient de ce coût par rapport à $\mathbf{D}$. On a ainsi pour $\forall\, p, q \in \{1...F\} \times \{1...K\}$ (le détail des calculs est donné en annexe A) :

$$\frac{\partial \mathscr{C}}{\partial \mathbf{D}_{pq}} = (\mathbf{W} \otimes (\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T))_{pq} - (\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T))_{pq} + \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{D}_{pq}}$$

Posons alors que :

$$\mathscr{C}_\mathbf{G}(\mathbf{G}) = \mathscr{C}_\mathbf{G}(\mathbf{D} \otimes \mathbf{W} - \mathbb{1}) = \mathscr{C}_\mathbf{D}(\mathbf{D})$$

On a classiquement pour $\forall\, p, q \in \{1...F\} \times \{1...K\}$:

$$\frac{\partial \mathscr{C}_\mathbf{D}}{\partial \mathbf{D}_{pq}} = \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{D}_{pq}} = \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{G}_{pq}} \frac{\partial \mathbf{G}_{pq}}{\partial \mathbf{D}_{pq}} = \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{G}_{pq}} \mathbf{W}_{pq}$$

soit encore sous forme matricielle :

$$\frac{\partial \mathscr{C}_\mathbf{D}}{\partial \mathbf{D}} = \frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{G}} \otimes \mathbf{W}$$

Posons en regroupant les termes positifs et négatifs du gradient du coût par rapport à $\mathbf{D}$:

$$\frac{\partial \mathscr{C}_\mathbf{D}}{\partial \mathbf{D}} = \mathbf{C}_\mathbf{D}^+ - \mathbf{C}_\mathbf{D}^-$$

et définisson $\mathbf{C}_\mathbf{G}^+$ et $\mathbf{C}_\mathbf{G}^-$ tels que :

$$\mathbf{C}_\mathbf{D}^+ = \mathbf{C}_\mathbf{G}^+ \otimes \mathbf{W} \qquad \text{et} \qquad \mathbf{C}_\mathbf{D}^- = \mathbf{C}_\mathbf{G}^- \otimes \mathbf{W}$$

ce qui implique d'après ce qui précède que :

$$\frac{\partial \mathscr{C}_\mathbf{G}}{\partial \mathbf{G}} = \mathbf{C}_\mathbf{G}^+ - \mathbf{C}_\mathbf{G}^-$$

Le gradient du coût total en fonction de $\mathbf{D}$ s'écrit donc également :

$$\frac{\partial \mathscr{C}}{\partial \mathbf{D}_{pq}} = (\mathbf{W} \otimes (\mathbf{V}^{\cdot(\beta-1)}\mathbf{H}^T))_{pq} - (\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{\cdot(\beta-2)})\mathbf{H}^T))_{pq} + \mathbf{C}^+_{\mathbf{D}\,pq} - \mathbf{C}^-_{\mathbf{D}\,pq}$$

A la différence de $\mathbf{G}$, les termes de $\mathbf{D}$ sont positifs comme nous avons vu précédemment : il est donc possible d'utiliser la mise à jour multiplicative classique pour $\mathbf{D}$ qui s'écrit donc à la $(n+1)^e$ iteration :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{C}^{-(n)}_{\mathbf{D}}}{\mathbf{W} \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{C}^{+(n)}_{\mathbf{D}}} \tag{33}$$

En appliquant à chaque terme de (33) un produit de Hadamard par $\mathbf{W}$ et en remplaçant $\mathbf{D}$ par son expression en fonction de $\mathbf{G}$ et les termes correspondant à $\mathscr{C}_{\mathbf{D}}$ par ceux correspondant à $\mathscr{C}_{\mathbf{G}}$, il vient :

$$\begin{aligned}
(33) &\iff \mathbf{D}^{(n+1)} \otimes \mathbf{W} = (\mathbf{D}^{(n)} \otimes \mathbf{W}) \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{W} \otimes \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{W} \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{W} \otimes \mathbf{C}^{+(n)}_{\mathbf{G}}} \\
&\iff \mathbf{G}^{(n+1)} + \mathbf{W} = (\mathbf{G}^{(n)} + \mathbf{W}) \otimes \frac{\mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{W} \otimes \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{W} \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{W} \otimes \mathbf{C}^{+(n)}_{\mathbf{G}}} \\
&\iff \mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} + \mathbf{W} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} - \mathbf{W} \\
&\iff \mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} + \mathbf{W} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T - \mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}} - \mathbf{C}^{+(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} \\
&\iff \mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} + \mathbf{W} \otimes \frac{((\mathbf{X} - \mathbf{V}^{(n)}) \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}} - \mathbf{C}^{+(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} \\
&\iff \mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}^{-(n)}_{\mathbf{G}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}^{+(n)}_{\mathbf{G}}} + \mathbf{\Gamma}^{(n)}
\end{aligned} \tag{34}$$

Annexe D Propriété de la déformation par tenseur

On considère un tenseur de K matrices $F \times F$. On s'intéresse à une matrice du tenseur, $\mathbf{D}$, et on considère le problème suivant :

$$\mathbf{y} = \mathbf{D}\mathbf{x} \tag{35}$$

$\mathbf{D}$ est une matrice de dimensions $F \times F$, $\mathbf{x}$ et $\mathbf{y}$ sont deux vecteurs de dimensions $F \times 1$. $\mathbf{x}$ et $\mathbf{y}$ sont connus et on cherche à exprimer $\mathbf{D}$. Il s'agit donc d'un problème sur-déterminé (et le déterminant d'un tel système est nul, il y a donc soit une infinité de solutions, soit aucune solution).

Or $\mathbf{y}\mathbf{x}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x} = \mathbf{y}\mathbf{x}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}$ puisque $\mathbf{x}$ est un vecteur et que $\mathbf{x}^T\mathbf{x}$ et son inverse sont donc des scalaires. On peut donc écrire $\mathbf{y} = \mathbf{y}\mathbf{x}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}$ et donc $\mathbf{D} = \mathbf{y}\mathbf{x}^T(\mathbf{x}^T\mathbf{x})^{-1} = \mathbf{y}\mathbf{x}^\dagger$ est une solution exacte particulière de (35).

On cherche à présent à trouver d'autres solutions en cherchant à estimer $\mathbf{D}$ minimisant une β-divergence $d_\beta(\mathbf{y}|\mathbf{D}\mathbf{x})$:

$$\begin{aligned}
\frac{\partial d_\beta(\mathbf{y}|\mathbf{D}\mathbf{x})}{\partial \mathbf{D}_{pq}} &= \sum_i (\mathbf{D}\mathbf{x})_i^{\beta-1}\frac{\partial(\mathbf{D}\mathbf{x})_i}{\partial \mathbf{D}_{pq}} - \mathbf{y}_i(\mathbf{D}\mathbf{x})_i^{\beta-2}\frac{\partial(\mathbf{D}\mathbf{x})_i}{\partial \mathbf{D}_{pq}} \\
&= \sum_i (\mathbf{D}\mathbf{x})_i^{\beta-2}\frac{\partial(\sum_j \mathbf{D}_{ij}\mathbf{x}_j)}{\partial \mathbf{D}_{pq}}((\mathbf{D}\mathbf{x})_i - \mathbf{y}_i) \\
&= \sum_{i,j} \delta_{ip}\delta_{jq}\mathbf{x}_j[(\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes (\mathbf{D}\mathbf{x}-\mathbf{y})]_i \\
&= [(\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes (\mathbf{D}\mathbf{x}-\mathbf{y})]_p\mathbf{x}_q \\
&= [((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes (\mathbf{D}\mathbf{x}-\mathbf{y}))\mathbf{x}^T]_{pq}
\end{aligned} \tag{36}$$

La formule de mise à jour sans contrainte s'écrit classiquement :

$$\mathbf{D} = \mathbf{D} \otimes \frac{((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes \mathbf{y})\mathbf{x}^T}{(\mathbf{D}\mathbf{x})^{\cdot(\beta-1)}\mathbf{x}^T}$$

On s'intéresse aux solutions obtenues si l'on soumet $\mathbf{D}$ à différentes contraintes que l'on note $\mathscr{C}(\mathbf{D})$

D.1 Cas $\mathscr{C}(\mathbf{D}) = \lambda\|\mathbf{D}\|_F^\alpha$

On impose à $\mathbf{D}$ de minimiser une puissance de sa norme de Frobenius. La dérivation de cette contrainte s'écrit :

$$\frac{\partial\mathscr{C}(\mathbf{D})}{\partial \mathbf{D}_{pq}} = \lambda\alpha\|\mathbf{D}\|_F^{\alpha-2}\mathbf{D}_{pq} \tag{37}$$

En combinant (36) et (37), le minimum est atteint pour $\mathbf{D}$ vérifiant :

$$((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes (\mathbf{D}\mathbf{x}-\mathbf{y}))\mathbf{x}^T + \lambda\alpha\|\mathbf{D}\|_F^{\alpha-2}\mathbf{D} = \mathbf{0} \tag{38}$$

La règle de mise à jour de $\mathbf{D}$ est donc :

$$\mathbf{D} = \mathbf{D} \otimes \frac{((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes \mathbf{y})\mathbf{x}^T}{((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes \mathbf{D}\mathbf{x})\mathbf{x}^T + \lambda\alpha\|\mathbf{D}\|_F^{\alpha-2}\mathbf{D}} \tag{39}$$

Cas particulier de la norme euclidienne

Dans le cas de la norme euclidienne $\beta = 2$, il existe une solution analytique pour (38) si $\alpha = 2$:

$$\mathbf{D} = \mathbf{y}\mathbf{x}^T(\mathbf{x}\mathbf{x}^T + 2\lambda\mathbf{I})^{-1} \tag{40}$$

L'introduction de la contrainte sur la norme de Frobenius permet de régulariser au sens de Tikhonov la matrice $\mathbf{x}\mathbf{x}^T$ qui est de rang 1, donc non inversible. Cette écriture revient donc à écrire $\mathbf{D} = \mathbf{y}\mathbf{x}^\dagger$.

D.2 Cas $\mathscr{C}(\mathbf{D}) = \lambda\|\mathbf{D}\|_1^\alpha$

On impose à $\mathbf{D}$ de minimiser une puissance de sa norme L_1. La dérivation de cette contrainte s'écrit :

$$\frac{\partial\mathscr{C}(\mathbf{D})}{\partial\mathbf{D}_{pq}} = \lambda\alpha\|\mathbf{D}\|_1^{\alpha-1} \tag{41}$$

En combinant (36) et (41), le minimum est atteint pour $\mathbf{D}$ vérifiant :

$$((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes (\mathbf{D}\mathbf{x} - \mathbf{y}))\mathbf{x}^T + \lambda\alpha\|\mathbf{D}\|_1^{\alpha-1}\mathbb{1} = \mathbf{0} \tag{42}$$

où $\mathbb{1}$ est la matrice composée de 1. La règle de mise à jour de $\mathbf{D}$ est donc :

$$\mathbf{D} = \mathbf{D} \otimes \frac{((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes \mathbf{y})\mathbf{x}^T}{((\mathbf{D}\mathbf{x})^{\cdot(\beta-2)} \otimes \mathbf{D}\mathbf{x})\mathbf{x}^T + \lambda\alpha\|\mathbf{D}\|_1^{\alpha-1}\mathbb{1}} \tag{43}$$

Cas particulier de la distance euclidienne

Dans le cas de la norme euclidienne $\beta = 2$, et en posant $\alpha = 1$, il existe une solution analytique puisque l'équation (42) devient :

$$\mathbf{D}\mathbf{x}\mathbf{x}^T = \mathbf{y}\mathbf{x}^T + \lambda\mathbb{1}$$

$\mathbf{x}\mathbf{x}^T$ étant de rang 1, son pseudo-inverse (qui est un scalaire) s'écrit $\mathbf{x}^\dagger = (\mathbf{x}^T\mathbf{x})^{-1}$ et

$$\mathbf{D} = (\mathbf{y}\mathbf{x}^T + \lambda\mathbb{1})\mathbf{x}^\dagger$$

Annexe E Correspondance de la NMF des déformation tensorielle et multiplicative terme à terme

Soient $\mathbf{W}_{cible}$ et $\mathbf{W}_{ref}$ 2 matrices de bases $F \times K$ dont on note les colonnes respectivement $\boldsymbol{\mathfrak{m}}_b$ et $\mathbf{w}_b$, $\forall b \in \{1...K\}$.

Soit $\boldsymbol{\Delta}$ un tenseur de K matrices $F \times F$ notées $\boldsymbol{\Delta}_b$, $\forall b \in \{1...K\}$, tel que :

$$\boldsymbol{\mathfrak{m}}_b = \boldsymbol{\Delta}_b \mathbf{w}_b = \mathbf{d}_b \otimes \mathbf{w}_b \qquad \text{avec} \qquad \mathbf{d}_b = (\boldsymbol{\Delta}_b \mathbf{w}_b)./\mathbf{w}_b \qquad \forall b \in \{1...K\} \tag{44}$$

En notant alors $\mathbf{W}_b$ et $\mathbf{D}_b$ les matrices $F \times K$ dont toutes les colonnes sont nulles sauf la b^e que l'on pose égale respectivement à $\mathbf{w}_b$ et $\mathbf{d}_b$, on peut écrire :

$$\mathbf{W}_{cible} = \boldsymbol{\Delta}\mathbf{W}_{ref} = \sum_b \boldsymbol{\Delta}_b \mathbf{W}_b = \sum_b \mathbf{D}_b \otimes \mathbf{W}_b = \mathbf{D} \otimes \mathbf{W}_{ref} \tag{45}$$

On rappelle que la règle de mise à jour sous contraintes du tenseur $\boldsymbol{\Delta}$ à la n^e iteration s'écrit comme la succession des mises à jour sous contraintes de chacune de ses K matrices $\boldsymbol{\Delta}_b$ suivant la règle :

$$\boldsymbol{\Delta}_b^{(n+1)} = \boldsymbol{\Delta}_b^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}}{\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}} \qquad \forall b \in \{1...K\} \tag{46}$$

Les termes $\mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}$ et $\mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}$ sont deux matrices $F \times F$ qui correspondent respectivement à la partie négative et à la partie positive du gradient des contraintes sur chaque matrice $\boldsymbol{\Delta}_b$ à la n^e itération.

On a donc en multipliant par le dénominateur :

$$\boldsymbol{\Delta}_b^{(n+1)} \otimes (\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}) = \boldsymbol{\Delta}_b^{(n)} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}) \qquad \forall b \in \{1...K\}$$

et en particulier :

$$(\boldsymbol{\Delta}_b^{(n+1)} \otimes (\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}))_{ik} = (\boldsymbol{\Delta}_b^{(n)} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}))_{ik} \qquad \forall b \in \{1...K\},\ \forall i,k \in [1...F] \tag{47}$$

ce qui *implique* que l'égalité suivante est également vraie :

$$\sum_{k=1}^{F} (\boldsymbol{\Delta}_b^{(n+1)} \otimes (\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}))_{ik} = \sum_{k=1}^{F} (\boldsymbol{\Delta}_b^{(n)} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T) + \mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)}))_{ik} \qquad \forall b \in \{1...K\},\ \forall i \in [1...F] \tag{48}$$

Or, on a pour 2 matrices carrées $\mathbf{A}$ et $\mathbf{B}$ l'égalité suivante :

$$\sum_{k=1}^{F} (\mathbf{A} \otimes \mathbf{B})_{ik} = \sum_{k=1}^{F} \mathbf{A}_{ik}\mathbf{B}_{ik} = (\mathbf{A}\mathbf{B}^T)_{ii}$$

Donc, pour $\forall b \in \{1...K\}, \forall i \in [1...F]$:

$$\begin{aligned}
(48) &\iff (\boldsymbol{\Delta}_b^{(n+1)}(\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + \mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)})^T)_{ii} = (\boldsymbol{\Delta}_b^{(n)}((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T) + \mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)})^T)_{ii} \\
&\iff (\boldsymbol{\Delta}_b^{(n+1)}\mathbf{W}_b\mathbf{H}(\mathbf{V}^{(n).(\beta-1)})^T + \boldsymbol{\Delta}_b^{(n+1)}\mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} = (\boldsymbol{\Delta}_b^{(n)}\mathbf{W}_b\mathbf{H}(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})^T + \boldsymbol{\Delta}_b^{(n)}\mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} \\
&\iff ((\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b)\mathbf{H}(\mathbf{V}^{(n).(\beta-1)})^T + \boldsymbol{\Delta}_b^{(n+1)}\mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} = ((\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b)\mathbf{H}(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})^T + \boldsymbol{\Delta}_b^{(n)}\mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} \\
&\iff ((\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b)\mathbf{H}(\mathbf{V}^{(n).(\beta-1)})^T)_{ii} + (\boldsymbol{\Delta}_b^{(n+1)}\mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} = ((\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b)\mathbf{H}(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})^T)_{ii} + (\boldsymbol{\Delta}_b^{(n)}\mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} \\
&\iff \sum_{k=1}^{F} ((\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b) \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T))_{ik} + (\boldsymbol{\Delta}_b^{(n+1)}\mathbf{C}^{+}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T})_{ii} = \sum_{k=1}^{F} ((\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b) \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T))_{ik} + (\boldsymbol{\Delta}_b^{(n)}\mathbf{C}^{-}_{\boldsymbol{\Delta}_\mathbf{b}}{}^{(n)T}
\end{aligned}$$

Or, par définition, $\mathbf{W}_{b,ik} = 0$ pour $k \neq b$ donc :

$$(48) \iff (\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T))_{ib} + (\mathbf{\Delta}_b^{(n+1)}\mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)T})_{ii} = (\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T))_{ib} + (\mathbf{\Delta}_b^{(n)}\mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)T})_{ii}$$

On pose alors $\mathbf{C}_{\mathbf{D}}^{+} = \sum_b \mathbf{C}_{\mathbf{D_b}}^{+}$ et $\mathbf{C}_{\mathbf{D}}^{-} = \sum_b \mathbf{C}_{\mathbf{D_b}}^{-}$ avec pour $\forall b, j \in [1...K], \forall i \in [1...F]$:

$$\begin{cases} \mathbf{C}_{\mathbf{D_b}\,ij}^{+\ (n)} & = \begin{cases} (\mathbf{\Delta}_b^{(n+1)}\mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)T})_{ii} / (\mathbf{\Delta}_b^{(n+1)}\mathbf{W}_b)_{ib} & \text{si } j = b \\ 0 & \text{sinon} \end{cases} \\ \mathbf{C}_{\mathbf{D_b}\,ij}^{-\ (n)} & = \begin{cases} (\mathbf{\Delta}_b^{(n)}\mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)T})_{ii} / (\mathbf{\Delta}_b^{(n)}\mathbf{W}_b)_{ib} & \text{si } j = b \\ 0 & \text{sinon} \end{cases} \end{cases} \tag{49}$$

On a alors, pour $\forall b \in \{1...K\}, \forall i \in [1...F]$:

$$\begin{aligned} (48) &\iff (\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T))_{ib} + (\mathbf{\Delta}_b^{(n+1)}\mathbf{W}_b)_{ib}\mathbf{C}_{\mathbf{D_b}\,ib}^{+\ (n)} = (\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T))_{ib} + (\mathbf{\Delta}_b^{(n)}\mathbf{W}_b)_{ib}\mathbf{C}_{\mathbf{D_b}\,ib}^{-\ (n)} \\ &\iff (\mathbf{D}_b^{(n+1)} \otimes \mathbf{W}_b \otimes (\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}_{\mathbf{D_b}}^{+\ (n)}))_{ib} = (\mathbf{D}_b^{(n)} \otimes \mathbf{W}_b \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}_{\mathbf{D_b}}^{-}))_{ib} \end{aligned}$$

soit encore sous forme matricielle :

$$\begin{aligned} (48) &\iff \mathbf{D}^{(n+1)} \otimes \mathbf{W} \otimes ((\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{+(n)}) = \mathbf{D}^{(n)} \otimes \mathbf{W} \otimes ((\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T) + \mathbf{C}_{\mathbf{D}}^{-(n)}) \\ &\iff \mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{-(n)}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{+(n)}} \end{aligned}$$

On reconnait là la règle de mise à jour avec contrainte d'une matrice de déformation terme à terme.

Il apparaît donc que la mise à jour multiplicative d'un tenseur de déformation est équivalente à la mise à jour multiplicative d'une matrice de déformation terme à terme construite selon (45).

On ne peut pas faire exactement le raisonnement inverse : en effet, (47) implique toujours (48), mais l'inverse n'est pas toujours vrai : il faudrait munir le tenseur de propriétés supplémentaires pour assurer l'équivalence. Une solution triviale est ainsi celle où tous les termes des sommes indexées par k sont nuls sauf celui pour $k = i$, ce qui correspond au cas où toutes les matrices du tenseur sont diagonales.

Il apparait également que si le tenseur a moins de matrices qu'il y a de bases, il n'est pas possible de passer d'une règle de mise à jour de $\mathbf{\Delta}$ à celle de $\mathbf{D}$ (la somme indexée par m ci-dessus ne peut pas disparaître), les deux déformations ne sont donc pas équivalentes.

Par ailleurs, on déduit de la première équation de (49) que pour $\forall b \in \{1...K\}, \forall i \in [1...F]$:

$$\begin{aligned} (49.1) &\iff (\mathbf{\Delta}_b^{(n+1)}\mathbf{W}_b)_{ib}\mathbf{C}_{\mathbf{D_b}\,ib}^{+\ (n)} = (\mathbf{\Delta}_b^{(n+1)}\mathbf{C}_{\mathbf{D_b}}^{+\ (n)T})_{ii} \\ &\iff \sum_{b'} (\mathbf{\Delta}_b^{(n+1)}\mathbf{W}_b)_{ib'}\mathbf{C}_{\mathbf{D_b}\,b'i}^{+\ (n)T} = (\mathbf{\Delta}_b^{(n+1)}\mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)T})_{ii} \qquad \text{(car les termes en b' sont nuls)} \\ &\iff (\mathbf{\Delta}_b^{(n+1)}\mathbf{W}_b\mathbf{C}_{\mathbf{D_b}}^{+\ (n)T})_{ii} = (\mathbf{\Delta}_b^{(n+1)}\mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)T})_{ii} \end{aligned}$$

On a de même pour $\forall b \in \{1...K\}, \forall i \in [1...F]$:

$$(49.2) \iff (\mathbf{\Delta}_b^{(n)}\mathbf{W}_b\mathbf{C}_{\mathbf{D_b}}^{-\ (n)T})_{ii} = (\mathbf{\Delta}_b^{(n)}\mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)T})_{ii}$$

On suppose $\mathbf{\Delta}_b$ inversible à chaque itération. Une condition *suffisante* pour que (49) soit vérifiée est donc que $\forall b \in \{1...K\}$:

$$\begin{cases} \mathbf{C}_{\mathbf{\Delta_b}}^{-\ (n)} & = \mathbf{C}_{\mathbf{D_b}}^{-\ (n)}\mathbf{W}_b^T \\ \mathbf{C}_{\mathbf{\Delta_b}}^{+\ (n)} & = \mathbf{C}_{\mathbf{D_b}}^{+\ (n)}\mathbf{W}_b^T \end{cases}$$

Comme $\forall b \in \{1...K\}$, $\mathbf{w}_b^T \mathbf{w}_b = \|\mathbf{w}_b\|_2^2$ est de dimension 1, il est inversible, et le système précédent est équivalent à, $\forall b \in \{1...K\}$:

$$\begin{cases} \mathbf{C}_{\mathbf{D_b}}^{-\,(n)} & = \mathbf{C}_{\mathbf{\Delta_b}}^{-\,(n)} \mathbf{W}_b / \|\mathbf{W}_b\|_2^2 \\ \mathbf{C}_{\mathbf{D_b}}^{+\,(n)} & = \mathbf{C}_{\mathbf{\Delta_b}}^{+\,(n)} \mathbf{W}_b / \|\mathbf{W}_b\|_2^2 \end{cases}$$

En notant $\mathscr{C}_{\mathbf{\Delta}_b}$ le coût que l'on applique à chaque matrice $\mathbf{\Delta}_b$ du tenseur et $\mathscr{C}_{\boldsymbol{D}}$ la contrainte que l'on applique terme à terme à $\mathbf{D}$, cette condition suffisante s'écrit :

$$\frac{\partial \mathscr{C}_{\boldsymbol{D}}}{\partial \boldsymbol{D}} = \sum_b \frac{\partial \mathscr{C}_{\mathbf{\Delta}_b}}{\partial \mathbf{\Delta}_b} \tilde{\mathbf{W}}_b \tag{50}$$

avec $\tilde{\mathbf{W}}_b = \mathbf{W}_b / \|\mathbf{W}_b\|_2^2$.

Cette relation reste valable à chaque itération de la mise à jour NMF. **Il apparaît donc que l'on peut toujours exprimer les contraintes imposées sur le tenseur sous la forme d'une contrainte équivalente sur la matrice de déformation terme à terme donnée par l'équation (50).**

E.1 Illustration de l'équivalence

Pour illustrer ce résultat, on envisage une déformation multiplicative par tenseur dont chacune des K matrices est soumise à une contrainte de "petite déformation", à savoir qu'elle doit minimiser une distance avec la matrice identité. On exprime cette contrainte sous la forme :

$$\mathscr{C}(\mathbf{\Delta}_b) = \|\mathbf{\Delta}_b - \mathbb{I}\|_2^2 \qquad \forall b \in \{1...K\}$$

On a alors simplement pour $\forall b \in \{1...K\}$:

$$\frac{\partial \mathscr{C}(\mathbf{\Delta}_b)}{\partial \mathbf{\Delta}_b} = 2(\mathbf{\Delta}_b - \mathbb{I})$$

Donc avec les notations du paragraphe précédent, on a $\mathbf{C}_{\mathbf{\Delta}}^{+\,(n)} = 2\mathbf{\Delta}_b^{(n)}$ et $\mathbf{C}_{\mathbf{\Delta}}^{-\,(n)} = 2\mathbb{I}$. La règle de mise à jour pour chaque matrice du tenseur de déformation s'écrit donc :

$$\mathbf{\Delta}_b^{(n+1)} = \mathbf{\Delta}_b^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})(\mathbf{W}_b\mathbf{H})^T + 2\mathbb{I}}{\mathbf{V}^{(n).(\beta-1)}(\mathbf{W}_b\mathbf{H})^T + 2\mathbf{\Delta}_b^{(n)}} \qquad \forall b \in [1...K]$$

En identifiant dans (13) les termes positifs et les termes négatifs, on calcule pour chaque n^e itération :

$$\mathbf{C}_{\mathbf{D}}^{+\,(n)} = 2\sum_b \mathbf{\Delta}_b^{(n)} \tilde{\mathbf{W}}_b = 2\boldsymbol{D}^{(n)} \otimes \tilde{\mathbf{W}}$$
$$\mathbf{C}_{\mathbf{D}}^{-\,(n)} = 2\sum_b \mathbb{I}\tilde{\mathbf{W}}_b = 2\tilde{\mathbf{W}}$$

On remarque alors que l'on peut intégrer le gradient des contraintes sur $\mathbf{D}$ et que l'on obtient :

$$\mathscr{C}_{\boldsymbol{D}} = \|(\mathbf{D} - \mathbb{I}) \otimes (\tilde{\mathbf{W}})^{\cdot(1/2)}\|_2^2$$

La contrainte d'identité sur le tenseur $\mathbf{\Delta}$ s'exprime sur la matrice $\mathbf{D}$ comme une contrainte d'identité, à un facteur multiplicatif près.

Note : dans le cas présent, l'apparition d'un terme en $\mathbf{\Delta W}$ qui est égal à $\mathbf{D} \otimes \mathbf{W}$ permet d'exprimer le gradient de la contrainte sur $\mathbf{D}$ simplement en fonction de $\mathbf{D}$, et donc de ne pas avoir à recourir au tenseur pour les calculs. Ce n'est cependant pas nécessairement le cas : une expression analytique des coûts sur $\mathbf{D}$ peut ne pas voir le terme en $\mathbf{\Delta}$ disparaître, ce qui réduit évidemment l'utilité de l'équivalence.

On calcule alors la mise à jour de $\mathbf{D}$ avec la formule équivalente :

$$\mathbf{D}^{(n+1)} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{-(n)}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + \mathbf{C}_{\mathbf{D}}^{+(n)}} = \mathbf{D}^{(n)} \otimes \frac{(\mathbf{X} \otimes \mathbf{V}^{(n).(\beta-2)})\mathbf{H}^T + 2\tilde{\mathbf{W}}}{\mathbf{V}^{(n).(\beta-1)}\mathbf{H}^T + 2\boldsymbol{D}^{(n)} \otimes \tilde{\mathbf{W}}}$$

On effectue alors une séparation pour un mélange voix + bruit de la base TIMIT avec un SNR de 0dB en utilisant les deux modèles de déformations mutliplicatives ($\mathbf{\Delta}$ et $\mathbf{D}$). On trace les valeurs des deux β-divergence pour β=0 à chaque itération (en rouge mise à jour avec le tenseur $\mathbf{\Delta}$, en bleu avec la matrice terme à terme $\mathbf{D}$, il y a bien deux courbes confondues) :

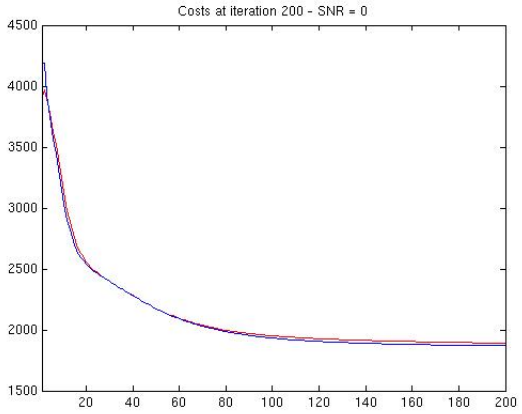


FIGURE 8 – Mise à jour avec SNR = 0dB

Cela confirme que les deux mises à jour sont équivalentes.

Interprétation Il est intéressant de noter que la contrainte sur $\mathbf{D}$ ne dépend d'aucune hypothèse faite sur la forme des matrices $\mathbf{\Delta}_b$. Par exemple, on aurait pu imposer aux matrices $\mathbf{\Delta}_b$ d'être diagonales, ou de n'avoir que certains coefficients non nuls autour de la diagonale, mais cette propriété ne serait pas apparue sur $\mathbf{D}$.

Il y a donc bien équivalence entre le cas sur-déterminé et le cas déterminé : pour cette contrainte, la sur-détermination du problème n'est pas pertinente pour que la NMF converge différemment que le cas déterminé.

Références

[Benaroya and Bimbot, 2003] Benaroya, L. and Bimbot, F. (2003). Wiener based source separation with hmm/gmm using a single sensor. In *Proc. ICA*, pages 957–961.

[Bouvier, 2015] Bouvier, D. (2015). Extraction de la voix dans un signal audio par factorisation en matrices non-négatives semi-supervisée. Rapport de master Atiam, Ircam.

[Dean et al., 2010] Dean, D. B., Sridharan, S., Vogt, R. J., and Mason, M. W. (2010). The qut-noise-timit corpus for the evaluation of voice activity detection algorithms. *Proceedings of Interspeech 2010*.

[Durrieu et al., 2009] Durrieu, J.-L., Ozerov, A., Févotte, C., Richard, G., and David, B. (2009). Main instrument separation from stereophonic audio signals using a source/filter model. In *Signal Processing Conference, 2009 17th European*, pages 15–19. IEEE.

[Févotte et al., 2009] Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence : With application to music analysis. *Neural computation*, 21(3) :793–830.

[Kitamura et al., 2013] Kitamura, D., Saruwatari, H., Shikano, K., Kondo, K., and Takahashi, Y. (2013). Music signal separation by supervised nonnegative matrix factorization with basis deformation. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–6. IEEE.

[Le Roux et al., 2015] Le Roux, J., Hershey, J. R., and Weninger, F. (2015). Deep nmf for speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 66–70. IEEE.

[Le Roux and Vincent, 2013] Le Roux, J. and Vincent, E. (2013). Consistent wiener filtering for audio source separation. *IEEE signal processing letters*, 20(3) :217–220.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

[Rix et al., 2001] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE.

[Smaragdis et al., 2007] Smaragdis, P., Raj, B., and Shashanka, M. (2007). Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Independent Component Analysis and Signal Separation*, pages 414–421. Springer.

[Souviraa-Labastie et al., 2015] Souviraa-Labastie, N., Olivero, A., Vincent, E., and Bimbot, F. (2015). Multi-channel audio source separation using multiple deformed references. *Audio, Speech, and Language Processing, IEEE/ACM Transactions*, pages 1775–1787.

[Sun and Mysore, 2013] Sun, D. L. and Mysore, G. J. (2013). Universal speech models for speaker independent single channel source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 141–145. IEEE.

[Villavicencio et al., 2006] Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE.

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

[Virtanen, 2003] Virtanen, T. (2003). Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, pages 35–40.

[Zue et al., 1990] Zue, V., Seneff, S., and Glass, J. (1990). Speech database development at mit : Timit and beyond. *Speech Communication*, 9(4) :351–356.