Un détecteur d'activité vocale visuel pour résoudre le problème des permutations en séparation de source de parole dans un mélange convolutif

Bertrand Rivet[1,2]*, Christine Servière*[2]*, Laurent Girin*[1]*, Dinh-Tuan Pham*[3]*, Christian Jutten*[2]

[1] Institut de la Communication Parlée (ICP)
CNRS UMR 5009, Institut National Polytechnique, Université Stendhal, Grenoble, France
[2] Laboratoire des Images et des Signaux (LIS)
CNRS UMR 5083, Institut National Polytechnique, Université Joseph Fourier, Grenoble, France
[3] Laboratoire de Modélisation et Calcul (LMC)
CNRS UMR 5523, Institut National Polytechnique, Université Joseph Fourier, Grenoble, France

ABSTRACT

Audio-visual speech source separation consists in mixing visual speech processing techniques (e.g. lip parameters tracking) with source separation methods to improve the extraction of a speech signal from a mixture of acoustic signals. In this paper, we present a new method that combines visual information with a separation method based on the sparseness of speech : visual information is used as a voice activity detector which is plugged on an acoustic separation technique. Results show the efficiency of the approach in the difficult case of realistic convolutive mixtures. Moreover, the overall process is quite simpler than previously proposed audiovisual separation schemes.

1. INTRODUCTION

La séparation de source aveugle consiste à retrouver des signaux sources à partir de mélanges de ces signaux, sans connaissances sur la nature du mélange ou sur les sources elles-mêmes. Pour les signaux de parole, la séparation n'est pas complètement aveugle car elle peut s'appuyer sur des propriétés spécifiques de ce signal. Par exemple, leur non-stationarité a été exploitée dans [4, 7]. Cependant, la séparation est encore une tâche difficile, notamment dans le cas où moins de capteurs que de sources sont disponibles, et aussi à cause des indéterminations de permutations et de gains : les signaux de sortie ne peuvent être correctement estimés qu'à un gain près et à une permutation près sur les canaux de sortie [1].

La séparation de source de parole audiovisuelle (SSAV) est un champ de recherche récent intéressant pour résoudre le problème de séparation dans le cas de signaux de parole [2, 8, 10]. Elle consiste à exploiter la bimodalité (audio/visuelle) de la parole pour améliorer les performances des systèmes de séparation acoustiques. En effet, les signaux visuels de la parole, en particulier les mouvements des lèvres du locuteur, fournissent une information complémentaire quand les signaux acoustiques sont dégradés par l'environnement. Sur cette base, Sodoyer *et al.* [8] puis Wang *et al.* [10] ont respectivement proposé d'utiliser un modèle statistique des cohérences entre traits visuels et acoustiques des signaux de parole pour extraire un signal de parole de mélanges de type additif et convolutif. Récemment, Rivet *et al.* [6] ont proposé une approche audiovisuelle similaire pour résoudre à la fois le problème du gain et des permutations dans le cas d'un mélange convolutif.

Dans cette nouvelle étude, on propose une approche différente pour résoudre le problème des permutations, potentiellement plus simple et plus efficace. L'information visuelle de la parole est utilisée comme un détecteur d'activité vocale (DAV) : son rôle est d'attester de la présence ou de l'absence du locuteur correspondant (celui qui est filmé) dans le mélange. Une telle information permet l'extraction du signal émis par ce locuteur.

Ce papier est organisé de la façon suivante. La Section 2 présente les bases du DAV visuel. La Section 3 rappelle les principes de la séparation de source pour des mélanges convolutifs et explique comment le DAV visuel peut être utilisé pour résoudre le problème des permutations pour le locuteur considéré. La Section 4 présente des résultats d'expérimentations.

2. DÉTECTEUR D'ACTIVITÉ VOCALE VISUEL

L'idée centrale du détecteur d'activité vocale visuel (DAV-V) est qu'en général, durant la production de parole, les lèvres bougent, alors qu'elles ne bougent pas (ou beaucoup moins) durant les silences. Nous utilisons le paramètre vidéo suivant :

$$v(m) = \left|\frac{\partial A(m)}{\partial m}\right| + \left|\frac{\partial B(m)}{\partial m}\right| \qquad (1)$$

où $A(m)$ et $B(m)$ sont les largeur et hauteur internes du contour labial. Ces paramètres sont extraits automatiquement toutes les 20ms (soit la longueur d'une *trame*) de façon synchrone à l'audio (échantillonné à 16kHz) en utilisant le système d'extraction développé à l'ICP [3]. La classification silence/parole est basée sur un seuillage. Cependant, le seuillage direct de $v(m)$ ne s'avère pas très performant : par exemple, les lèvres peuvent être immobiles pendant plusieurs trames, alors que le locuteur est en train de parler. C'est pourquoi, $v(m)$ est d'abord lissé par intégration temporelle sur T trames consécutives : $V(m) = \sum_{l=0}^{T-1} a_l\, v(m-l)$ où les a_l sont les coefficients d'un filtre passe-bas IIR du premier ordre. La trame m est alors classifiée comme silence si $V(m)$ est inférieure à un seuil δ et elle est classifiée comme parole sinon. Comme expliqué à la Section 3, l'objectif du DAV-V est en fait de détecter les trames de signal où le locuteur filmé *ne produit pas de son*. Pour diminuer le taux de fausses alarmes (décision silence pendant l'activité de parole), seules les séquences d'au moins L trames de silence sont finalement considérées comme silence. Au final, le DAV-V proposé est robuste à n'importe quel bruit environnant et peut être exploité même dans un environnement sonore hautement non-stationnaire, quels que soient le nombre et la nature des sources concurrentes. On peut trouver plus de détails dans [9].

3. SÉPARATION DE SOURCE AVEC DAV-V

Dans cette section, on présente d'abord brièvement le cadre général de la séparation de sources pour des mélanges convolutifs stationaires, puis nous expliquons comment le DAV-V peut être utilisé pour résoudre le problème des permutations.

3.1. Séparation de source de mélange convolutif

Considérons le cas général de N sources $\mathbf{s}(m) = [s_1(m), \cdots, s_N(m)]^T$ à extraire à partir de P observations $\mathbf{x}(m) = [x_1(m), \cdots, x_P(m)]^T$ (^{T} dénote la transposition) : $x_p(m) = \sum_{n=1}^{N} h_{p,n}(m) * s_n(m)$. Les filtres $h_{p,n}(m)$ modélisant la réponse impulsionnelle entre chaque source $s_n(m)$ et le $p^{ème}$ capteur, sont les éléments de la matrice de mélange $H(m)$. Le but de la séparation est de récupérer les sources par un filtrage dual : $\hat{s}_n(m) = \sum_{p=1}^{P} g_{n,p}(m) * x_p(m)$ où les $g_{n,p}(m)$ sont les éléments de la matrice de séparation $G(m)$ et sont estimés de façon à ce que les sources estimées en sortie $\hat{\mathbf{s}}(m) = [\hat{s}_1(m), \cdots, \hat{s}_N(m)]^T$ soient les plus indépendantes possibles (ou au moins décorrélées) deux à deux. Ce problème est généralement traité dans le domaine fréquentiel, par exemple [4, 7], on a alors :

$$X_p(m,f) = \sum_{n=1}^{N} H_{p,n}(f) S_n(m,f) \quad (2)$$

$$\hat{S}_n(m,f) = \sum_{p=1}^{P} G_{n,p}(f) X_p(m,f) \quad (3)$$

où $S_n(m,f)$, $X_p(m,f)$ et $\hat{S}_n(m,f)$ sont respectivement les transformées de Fourier à court terme (TFCT) de $s_n(m)$, $x_p(m)$ et $\hat{s}_n(m)$. $H_{p,n}(f)$ et $G_{n,p}(f)$ sont respectivement les réponses en fréquence des filtres de mélange et de séparation. Des manipulations algébriques simples sur (2) et (3) conduisent à :

$$\Gamma_x(m,f) = H(f)\Gamma_s(m,f)H^H(f) \quad (4)$$

$$\Gamma_{\hat{s}}(m,f) = G(f)\Gamma_x(m,f)G^H(f) \quad (5)$$

où $\Gamma_y(m,f)$ dénote la matrice de densité spectrale de puissance (DSP) à court terme d'un signal multidimensionnel $\mathbf{y}(m)$. $H(f)$ et $G(f)$ sont respectivement les matrices de réponse en fréquence associées aux matrices de mélange et de séparation (^{H} dénote le transposé conjugué). Si on suppose que les sources sont mutuellement indépendantes (ou au moins décorrélées), $\Gamma_s(m,f)$ est diagonale et une séparation efficace doit conduire à une matrice diagonale $\Gamma_{\hat{s}}(m,f)$. Par conséquent, un critère basique pour la séparation est de calculer $\Gamma_x(m,f)$ à partir des observations et d'ajuster la matrice $G(f)$ de telle façon que $\Gamma_{\hat{s}}(m,f)$ soit aussi diagonale que possible. Comme cette condition doit être vérifiée pour n'importe quel indice temporel m, ceci peut être fait par un algorithme de diagonalisation conjointe (*i.e.* la meilleure diagonalisation simultanée de plusieurs matrices) [5], et par la suite nous utilisons l'algorithme de séparation par diagonalisation conjointe des matrices de DSP de Servière et Pham [7].

3.2. Résolution du problème des permutations

La limitation classique des algorithmes de séparation est que pour chaque canal fréquentiel, $G(f)$ ne peut être estimée qu'à un gain près et à une permutation près entre les sources : $G(f) = P(f)\,D(f)\,\hat{H}^{-1}(f)$ où $P(f)$ et $D(f)$ sont une matrice de permutation et une matrice diagonale arbitraires. Plusieurs approches purement audio au problème des permutations ont été proposées (par exemple [4, 7]). Dans [6], nous avons proposé d'utiliser un modèle statistique des cohérences audiovisuelle des signaux de parole pour lever les indéterminations de permutation et de gain . Bien qu'efficace, la méthode à les désavantages de nécessiter un apprentissage hors-ligne et d'être coûteuse en calcul.

Dans cette nouvelle étude, nous simplifions cette approche en exploitant l'information plus simple délivrée par le DAV-V focalisant sur les lèvres du locuteur dont on veut extraire le signal de parole. Le modèle audiovisuel de [6] est remplacée par le DAV-V de la Section 2 et la détection de l'absence de la source d'intérêt permet de régulariser le problème de permutation pour cette source. En effet, pour chaque fréquence f, l'algorithme de séparation fournit une matrice de séparation $G(f)$ qui conduit à une matrice de DSP des sources estimées $\Gamma_{\hat{s}}(m,f)$ diagonale. Le $k^{ème}$ élément de la diagonale de $\Gamma_{\hat{s}}(m,f)$ représente la variation de l'énergie spectrale de la $k^{ème}$ source estimée à la fréquence f au cours du temps m. Appelons le logarithme de cette valeur un *profil* et notons-le $E(f,m;k)$. Notons $\mathcal{T}$ l'ensemble de tous les indices temporels. Supposons maintenant qu'un DAV-V, associé à une source particulière, disons $s_1(m)$, nous fournit l'ensemble des indices temporels $\mathcal{T}_1$ où cette source disparaît du mélange ($\mathcal{T}_1 \subset \mathcal{T}$). Alors le profil $E(f,m;\cdot)$, avec $m \in \mathcal{T}_1$, correspondant à l'estimation de $s_1(m)$ doit être proche de -∞. Par conséquent, nous proposons la technique de régularisation de permutation suivante pour extraire la source particulière $s_1(m)$ du mélange $\mathbf{x}(m)$. A la sortie de l'algorithme de diagonalisation conjointe, on calcule les profils centrés $E_{\mathcal{T}_1}(f;k)$ pendant que $s_1(m)$ est détectée absente, soit pendant $m \in \mathcal{T}_1$:

$$E_{\mathcal{T}_1}(f;k) = \frac{1}{|\mathcal{T}_1|}\sum_{m\in\mathcal{T}_1} E(f,m;k) - \frac{1}{|\mathcal{T}|}\sum_{m\in\mathcal{T}} E(f,m;k) \quad (6)$$

où $|\mathcal{T}_1|$ est le cardinal de l'ensemble $\mathcal{T}_1$. Le centrage permet d'éliminer l'influence du gain non contrôlé, puisque celui-ci devient une constante additive en échelle log. Puis, à partir du fait que le profil centré $E_{\mathcal{T}_1}(f;\cdot)$ correspondant à l'estimation de $s_1(m)$ doit être proche de -∞, pour toutes les fréquences f, on recherche le profil centré de plus faible valeur. On règle alors $P(f)$ de façon à ce que cette valeur minimale corresponde à $E_{\mathcal{T}_1}(f;1)$. L'application de cet ensemble de matrices de permutation $P(f)$ aux matrices de séparation $G(f)$ permet de reconstruire la source $s_1(m)$ sans permutations en fréquence.

Notons que la méthode proposée permet de résoudre les permutations de fréquence pour la source à laquelle le DAV-V est associé, mais il peut rester des permutations entre les autres sources sans conséquence pour l'extraction de $s_1(m)$. Pour extraire plus d'une source, il est nécessaire d'avoir des DAV-V supplémentaires correspondant.

4. EXPÉRIMENTATIONS

Dans cette section, on considère le cas de deux sources mélangées par des matrices de filtres 2×2. Ces filtres sont des filtres RIF de 512 coefficients avec trois échos principaux. Ils sont extraits d'une librairie de réponses impulsionnelles mesurées dans une grande pièce de 3.5m×7m×3m (on peut les trouver

à `http ://sound.medi.mit.edu/ica-bench`). Le corpus utilisé pour la source $s_1(t)$ à extraire est de la parole continue produite par un locuteur masculin enregistré en condition de dialogue spontané. La seconde source est de la parole continue produite par un autre locuteur.

Pour caractériser les performances de l'extraction de $s_1(t)$, on définit un indice de performance :

$$r_1(f) = |GH_{12}(f)/GH_{11}(f)|\,, \qquad (7)$$

où $GH_{i,j}(f)$ est le $(i,j)^{ème}$ élément du système global $GH(f) = G(f)\,H(f)$. Pour une bonne séparation, cet indice doit être proche de 0 (∞ si une permutation a eu lieu).

La Figure 1 présente un exemple de séparation. Les Figures 1(a) et 1(b) montrent les deux sources et les mélanges. Dans ces expériences, dix secondes de signal sont utilisées pour estimer des filtres de séparation de 4096 coefficients (ce qui est donc la taille de toutes les TFCT). Les Figures 1(c), 1(e) et 1(g) montrent les sources estimées dans différentes conditions (voir ci-dessous). Les indices $r_1(f)$ correspondants, tronqués à 1, sont représentés sur les Figures 1(d), 1(f) et 1(h). Sur les Figures 1(a), 1(c), 1(e) et 1(g) les pointillés représentent une indexation manuelle des silences et les traits discontinus représentent la détection automatique obtenue avec le DAV-V de la Section 2. On peut voir que cette détection est performante. De résultats plus détaillés sont donnés dans [9].

Dans la première expérience (Figures 1(c) et 1(d)), les sources sont estimées par l'algorithme de diagonalisation conjointe sans régularisation des permutations. On peut voir que plusieurs blocs de fréquences consécutives sont permutés, ainsi que plusieurs fréquences isolées (Figure 1(d)). Par conséquent, les signaux séparés contiennent des composantes basses/hautes fréquences permutées entre les deux sources (Figure 1(c)). Dans la deuxième expérience (Figures 1(e) et 1(f)), les sources sont estimées par l'algorithme de diagonalisation conjointe utilisant en plus la variation relative des profils [7] : les permutations sont détectées en se basant sur le fait que les profils $E(f,m;k)$ d'une source donnée varient de façon lisse avec la fréquence. On peut voir que les permutations principales sont corrigées, ce qui permet une bonne estimation des sources. Cependant, plusieurs permutations isolées restent présentes bien qu'elles aient une influence limitée sur la qualité de la séparation : sur la Figure 1(f), $(G * H)_{1,1}(n)$ est largement supérieur à $(G * H)_{1,2}(n)$. Dans la dernière expérience (Figure 1(g) et 1(h)), les sources sont estimées en utilisant le DAV-V du locuteur 1 pour détecter les silences de $s_1(m)$ et ainsi régulariser les permutations en utilisant la technique de la Section 3.2. A partir des résultats de [9], on a choisi $\alpha_l = 0.82^l$ et le nombre minimal de trames de silence consécutives L est de 20 (*i.e.* la longueur minimum d'un silence est de 400ms). On peut voir que la méthode proposée permet une très bonne estimation des sources. Il reste quelques permutations isolées mais une investigation plus profonde révèle qu'elles correspondent à des régions des spectres avec une très faible énergie pour les deux sources : elles ont donc une influence très faible sur la qualité de la séparation, comme on peut le voir à la Figure 1(h). Les Figures 1(i) et 1(j) montrent les profils centrés des deux sources estimées avant et après la correction de permutations par le DAV-V. On voit que les blocs permutés sont bien détectés par les profils calculés à partir de la détection de silence : après la régularisation de permutations, on a $E_{\mathcal{T}_1}(f;1) \leq E_{\mathcal{T}_1}(f;2)$ ce qui conduit à une bonne estimation des sources. Ces observations sont confirmées par l'écoute des signaux.

5. CONCLUSION

La détection visuelle, robuste à tout type d'environnement sonore, s'est révélée efficace pour régulariser le problème des permutations selon un principe très simple. Notons aussi que cette méthode visuelle a un avantage important par rapport aux méthodes de séparation purement audio [7] qui fournissent les sources dans un ordre arbitraire (*i.e.* à une permutation globale près sur les sortie même si les permutations sur les différentes fréquences sont corrigées) : l'information visuelle permet d'associer un canal de sortie au locuteur filmé. Dans ce travail, toutes les détections sont réalisées hors-ligne, c'est-à-dire sur des sections de signal relativement larges (de l'ordre de 10s). Nos travaux futurs concerneront le développement d'une version pseudo-temps-réel où les traitements sont effectués en ligne, pour se rapprocher des conditions d'utilisation réelles.

RÉFÉRENCES

[1] Jean-François Cardoso. Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10) :2009–2025, October 1998.

[2] R.M. Dansereau. Co-channel audiovisual speech separation using spectral matching constraints. In *Proc. ICASSP*, Montréal, Canada, 2004.

[3] T. Lallouache. Un poste visage-parole. Acquisition et traitement des contours labiaux. In *Proc. Journées d'Etude sur la Parole (JEP) (French)*, Montréal, 1990.

[4] Lucas Para and Clay Spence. Convolutive blind separation of non stationary sources. *IEEE Trans. Speech Audio Processing*, 8(3) :320–327, May 2000.

[5] Dinh-Tuan Pham. Joint approximate diagonalization of positive definite matrices. *SIAM J. Matrix Anal. And Appl.*, 22(4) :1136–1152, 2001.

[6] Bertrand Rivet, Laurent Girin, and Christian Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Trans. Speech Audio Processing*, (Accepted for publication).

[7] Christine Servière and Dinh-Tuan Pham. A novel method for permutation correction in frequency-domain in blind separation of speech mixtures. In *Proc. ICA*, pages 807–815, Granada, Spain, 2004.

[8] David Sodoyer, Laurent Girin, Christian Jutten, and Jean-Luc Schwartz. Developing an audio-visual speech source separation algorithm. *Speech Comm.*, 44(1–4) :113–125, October 2004.

[9] David Sodoyer, Bertrand Rivet, Laurent Girin, Jean-Luc Schwartz, and Christian Jutten. An analysis of visual speech information applied to voice activity detection. In *Proc. ICASSP*, Toulouse, France, 2006 (accepted).

[10] Wenwu Wang, Darren Cosker, Yulia Hicks, Saied Sanei, and Jonathon A. Chambers. Video assisted speech source separation. In *Proc. ICASSP*, Philadelphia, USA, March 2005.

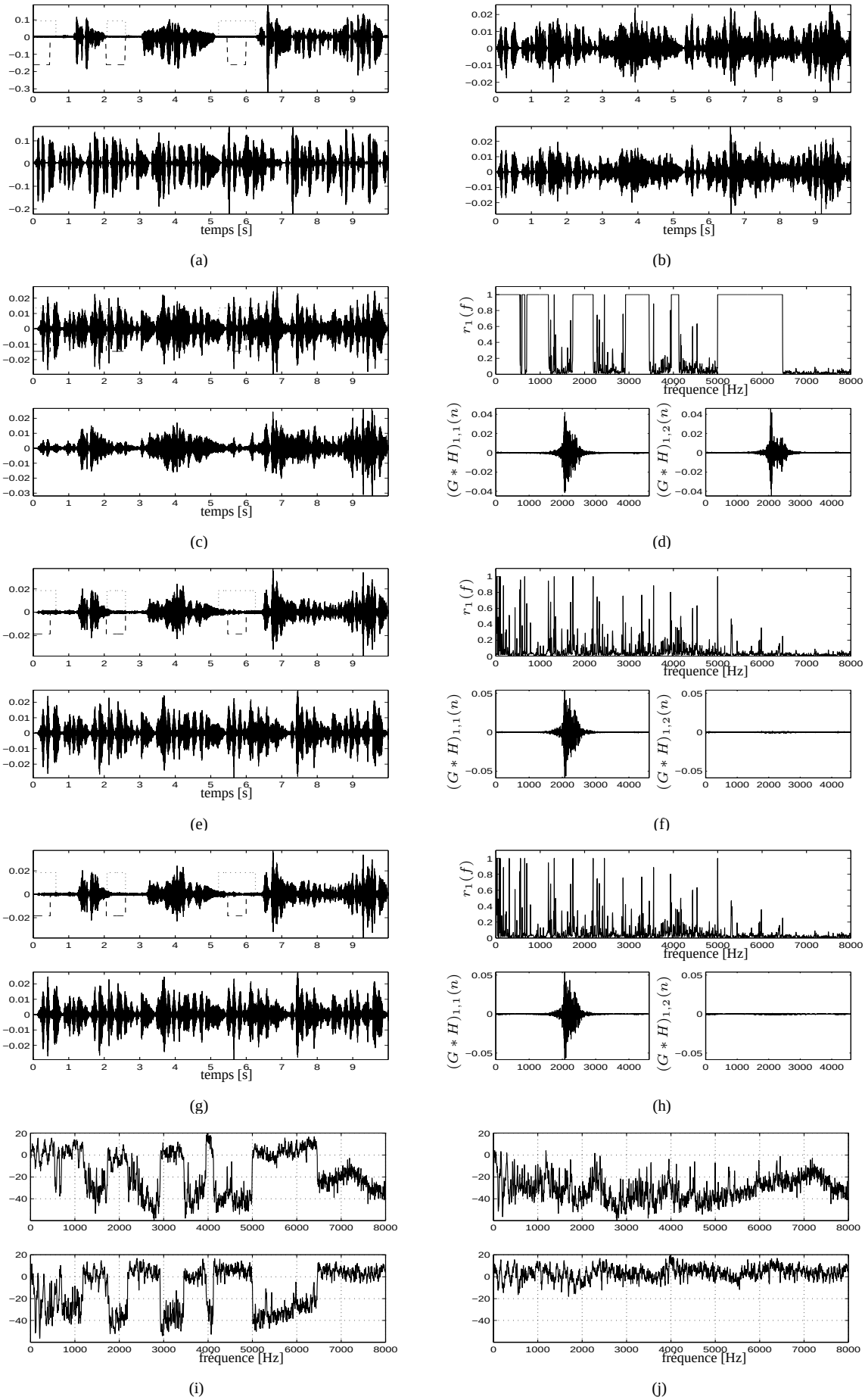

FIG. 1: Sources (a), mélanges (b), sources estimées (c), (e), (g), indice de performance (tronqués à 1) et réponses impulsionnelles du système global $(G*H)(n)$ (d), (f), (h) et profils centrés avant (i) et après (j) corrections des permutations.