TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

SEPARATION AVEUGLE DE PAROLE ET DE BRUIT DANS UN MELANGE CONVOLUTIF

H.L. Nguyen Thi*, Ch. Jutten**, J. Caelen*

* ICP/INPG **TIRF/INPG
46 Av. Félix Viallet, 38031 Grenoble Cedex

RÉSUMÉ

Un algorithme de séparation parole-bruit à partir d'un enregistrement stéréophonique est présenté dans cet article. A la sortie de chaque microphone, on obtient un signal composite résultant du mélange convolutif inconnu du signal parole et du bruit ambiant (modélisé par des filtres RIF inconnus). L'algorithme mis en œuvre est une généralisation du modèle additif de séparation aveugle de sources de Hérault-Jutten (H-J) pour le cas du mélange convolutif. La séparation peut être obtenue par une estimation simultanée des filtres inverses dont les coefficients sont mis à jour par un algorithme adaptatif. Les résultats de simulation sont commentés, ils confirment l'intérêt de cette approche (gain de 19 dB en moyenne).

ABSTRACT

An algorithm for the separation of the speech signal from noise in stereophonic speech recording with two microphones is described. In this situation, the signal measured at the output of each microphone is a superimposition of two unknown sources in the convolutive mixture which is modeled by unknown linear filters (FIR). This algorithm is a generalization of the Hérault-Jutten (H-J) algorithm for the blind separation of sources in the case of a convolutive mixture. Thus, the separation of the speech signal from noise can be achieved by estimating simultaneously the inverse filters in which the coefficients are updated by an adaptive algorithm. Experimental results (gain of 19 dB on average) are discussed.

Introduction

Le problème de séparation du signal et du bruit reste encore un des problèmes importants en traitement de la parole. Un signal de parole enregistré dans un environnement réel est normalement dégradé par des bruits ambiants qui limitent les performances des systèmes de reconnaissance automatique de la parole. Il est donc nécessaire de rehausser le signal de parole le plus possible avant de le fournir à un système de reconnaissance. De nombreux algorithmes de rehaussement du signal ont été proposés : il semble que les performances des systèmes de suppression de bruit d'interférence avec un microphone soient limitées [3], en particulier dans les cas où on ne connaît ni les signaux primitifs ni la loi du mélange des signaux. Une des solutions peut consister à capter la parole, à l'aide de deux microphones, puis à séparer le signal et le bruit primitifs (sources). Le modèle de H-J est une solution neuromimétique à un problème de séparation aveugle de sources. Nous présentons dans cet article une solution fondée sur une généralisation de ce modèle[1] — présenté initialement pour des mélanges additifs — au cas du mélange convolutif. La solution peut être obtenue sans hypothèses particulières sur des signaux : stationnaires ou non, à bande large ou à bande étroite, déterministes ou aléatoires, etc.
Cet article est organisé en quatre parties. La première définit le modèle de perturbation basé sur les filtres de type FIR. La seconde développe un algorithme généralisé de celui de H-J dans le cas de mélange convolutif pour séparer le signal de la parole et le bruit. Les résultats de simulation sont présentés dans la troisième partie. Des discussions et conclusions constituent la dernière partie.

1. Modèle de mélange

Dans le cas d'une prise de son par deux microphones dans un environnement réel (salle de saisie bureautique), on obtient, à la sortie d'un microphone, un signal complexe, résultant de la superposition du signal de parole et des bruits ambiants selon un mélange inconnu. Ce mélange dépend des positions des microphones, des caractéristiques acoustiques de la salle, des sources elles-mêmes, etc. En général, un tel mélange est un mélange convolutif dans lequel les signaux ont les propriétés suivantes :
- le signal de la parole a une bande passante large,
- le bruit est généralement non stationnaire, non ponctuel, non blanc, et peut avoir un niveau élevé.

Approximativement on peut considérer qu'un mélange convolutif résulte du passage des signaux dans des filtres inconnus de type FIR [3]. Le modèle général de mélange est présenté dans la figure 1, où A, B, C, D sont les systèmes linéaires représentant les fonctions de transfert entre les sources et les deux microphones. Ce modèle général de mélange se complique par le fait que les sources X_i et tous les filtres sont inconnus.

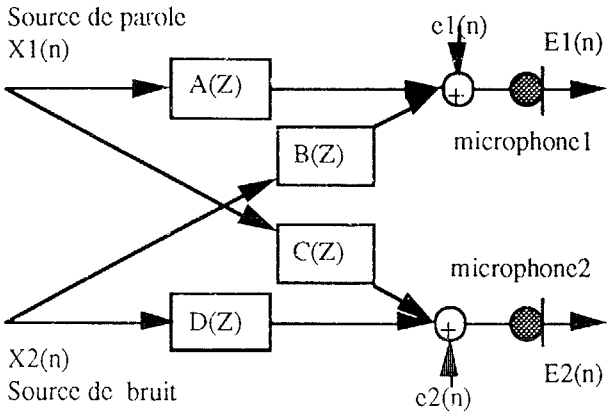

Fig. 1. Modèle général de mélange convolutif.

Dans une première approximation, nous supposons que :
- les sources sont quasiment ponctuelles,
- un microphone est placé judicieusement près du locuteur et l'autre près de la source de bruit,
- la réponse en fréquence des microphones est presque constante sur tout le spectre de parole,
- la distance entre les deux microphones est petite,
- les erreurs de mesure des microphones $e_1(n)$ et $e_2(n)$ sont négligeables.

Dans ces conditions nous pouvons prendre un modèle simplifié de perturbation [3], [4] dans lequel les deux filtres A(Z) et D(Z) sont des "passe-tout" (Fig. 2).

2. Modèle de séparation

Le principe du modèle de séparation de sources est d'extraire simultanément dans les signaux de mélange inconnu, tous les signaux primitifs. Dans le cas d'un mélange additif, Hérault et Jutten ont proposé un algorithme adaptatif, dont le principe repose sur un test d'indépendance, obtenu en introduisant des fonctions non-linéaires dans une règle d'adaptation proche de celle de Widrow-Hoff. Dans le cas du mélange convolutif, les deux coefficients (scalaires) adaptatifs du modèle de H-J deviennent deux filtres adaptatifs.

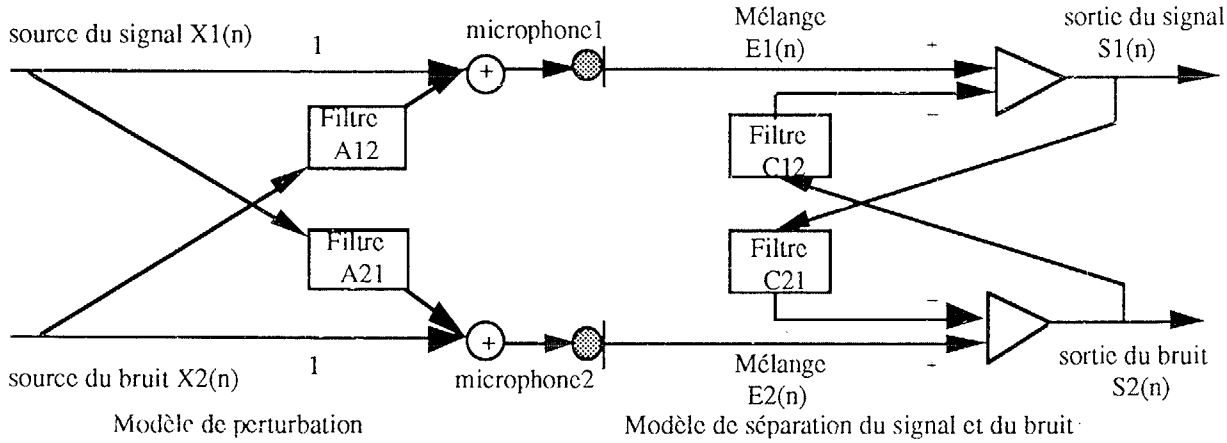

Fig. 2. Schéma synoptique du modèle de séparation du signal et du bruit.

2.1 Architecture

En utilisant les hypothèses précédentes, à chaque instant n, on aboutit alors aux équations de mélange convolutif :

$$E_i(n) = X_i(n) + A_{ij}(n) * X_j(n) \qquad (1)$$

avec $i,j = [1,2]$ et $i \neq j$.

soit en développant l'expression des filtres FIR :

$$E_i(n) = X_i(n) + \sum_{k=0}^{M-1} a_{ij}(k).X_j(n-k) \qquad (1')$$

où les deux signaux primitifs $X_1(n)$ et $X_2(n)$ sont indépendants et inconnus, les deux filtres A_{12} et A_{21} sont également inconnus et d'ordre M.

Dans ce cas, les équations à la sortie du modèle de séparation de sources de H-J sont :

$$S_i(n) = E_i(n) - \sum_{k=0}^{M-1} c_{ij}(k).S_j(n-k) \qquad (2)$$

où C_{ij} sont les filtres adaptatifs estimés.

Prenons la transformée en Z des équations (1) et (2). En résolvant le système on aboutit à :

$$S_1(z) = \frac{(1 - C_{12}.A_{21}).X_1(z) + (A_{12} - C_{12}).X_2(z)}{(1 - C_{12}.C_{21})}$$

$$S_2(z) = \frac{(1 - C_{21}.A_{12}).X_2(z) + (A_{21} - C_{21}).X_1(z)}{(1 - C_{12}.C_{21})} \qquad (3)$$

En considérant les équations (3), on peut arriver à une solution de séparation de sources si l'on impose :

$$C_{12}(z) = A_{12}(z) \quad \text{et} \quad C_{21}(z) = A_{21}(z) \qquad (4)$$

qui rend $S_1(z)$ proportionnel à $X_1(z)$ et $S_2(z)$ proportionnel à $X_2(z)$, ce qui correspont à une solution de séparation de $S_1(n)$ et de $S_2(n)$ dans le domaine temporel.

2.2 Algorithme

Les deux filtres A_{12} et A_{21} du mélange sont inconnus. Par conséquent, les deux filtres C_{12} et C_{21} ne peuvent qu'être estimés par un algorithme (récursif) de telle sorte qu'ils s'approchent des deux filtres inconnus A_{ij} à la convergence de l'algorithme.

Considérons alors la sortie $S_i(n)$ comme un terme d'erreur de l'estimation linéaire $\hat{E}_i(n)$ de l'entrée $E_i(n)$ à partir de l'autre sortie $S_j(n)$ [1] :

$$S_i(n) = E_i(n) - \sum_{k=0}^{M-1} c_{ij}(k).S_j(n-k) = E_i(n) - \hat{E}_i(n) \qquad (5)$$

En calculant les dérivées partielles de la moyenne de $S_i^2(n)$ par

rapport à tous les coefficients de filtres $c_{ij}(k)$:

$$\frac{\partial E\left[S_i^2(n)\right]}{\partial c_{ij}^n(k)} = -2E\left[S_i(n).S_j(n-k)\right] \quad (6)$$

et en généralisant la règle d'adaptation dissymétrique du modèle de H-J — à chaque coefficient $c_{ij}(k)$ des filtres — à l'instant n+1, on aboutit à la règle adaptative :

$$c_{ij}^{n+1}(k) = c_{ij}^n(k) + \mu.f(s_i(n)).g(s_j(n-k)) \quad (7)$$

où :

$s_i(n)$ est une estimation du signal centré de $S_i(n)$,

μ est un gain d'adaptation positif,

$k \in [0, M-1]$, M étant l'ordre des filtres,

les deux fonctions f et g sont non-linéaires [1], [7].

Puis, en choisissant une formulation simple pour les fonctions f et g :

$$f(.) = (.)^3 \text{ et } g(.) = (.)$$

on déduit la règle finale suivante pour estimer les coefficients des filtres du modèle :

$$c_{ij}^{n+1}(k) = c_{ij}^n(k)\, n + \mu.\, s_i^3(n).\, s_j(n-k) \quad (8)$$

2.3 La convergence

On peut considérer le principe de cet algorithme comme voisin de celui de l'algorithme LMS [2] dans lequel les filtres adaptatifs sont estimés par la formule récursive :

$$C_{ij}(n+1) = C_{ij}(n) + \mu\nabla_i(n) \quad (9)$$

où $\nabla_i(n)$ est un vecteur du gradient de filtre.

Dans ce cas, on peut donc écrire que :

$$\frac{\partial J_i(n)}{\partial C_{ij}(n)} = -2\left[s_i(n).S_{sj}(n)\right] = \nabla_i(n) \quad (10)$$

où S_{sj} est le vecteur des sorties :

$$S_{sj}(n) = \begin{bmatrix} s_j(n) \\ s_j(n-1) \\ \cdots \\ s_j(n-k) \\ \cdots \\ s_j(n-M-1) \end{bmatrix}$$

Pour i = 1 et 2, on a :

$$E\left[\nabla_1(n)\right] = -2E\left[s_1(n).S_{s2}(n)\right]$$
$$= -2E\left[\left(e_1(n) - S_{s2}^T(n).C_{12}(n)\right).S_{s2}(n)\right]$$
$$= 2(R_2.C_{12} - P_1)$$
$$E\left[\nabla_2(n)\right] = -2E\left[s_2(n).S_{s1}(n)\right] = 2(R_1.C_{21} - P_2)$$

En remplaçant les indices muets, on obtient :

$$E\left[\nabla_i(n)\right] = 2(R_j.C_{ij} - P_i) \quad (11)$$

où :

$S_{sj}^T(n)$ est un vecteur transposé du vecteur $S_{sj}(n)$,

$$R_i = E\left[S_{si}(n).S_{si}^T(n)\right] \quad (12)$$

est la matrice de covariance de la sortie $S_i(n)$,

$$P_i = E\left[e_i(n).S_{sj}(n)\right] \quad (13)$$

est le vecteur d'intercorrélation de l'entrée i et du vecteur de sortie j.

A la convergence de l'algorithme (voisin de l'algorithme LMS), on obient alors :

$$E\left[s_i(n).S_{sj}(n)\right] = 0 \quad (14)$$

qui signifie que la covariance d'une sortie et de l'autre retardée de k échantillons, est nulle.

Les deux filtres C_{ij} convergent vers la solution de l'équation de Wiener-Hoff. Dans ce cas, la solution est :

$$C_{12}^* = R_2^{-1}.P_1 \quad \text{et} \quad C_{21}^* = R_1^{-1}.P_2 \quad (15)$$

Ces relations garantissent la non-corrélation des sorties mais non l'indépendance statistique. Dans la règle utilisée, les fonctions non-linéaires f et g introduisent des moments d'ordre supérieur et les filtres estimés tendent vers une solution réalisant la séparation des sources. :

$$E\left[s_i^3(n).S_{sj}(n)\right] = 0 \quad (16)$$

Cette relation comprend un test d'indépendance des sorties, sauf dans des cas très particuliers de densité de probabilié de sources inconnues [7], [8], et on obtient alors :

$$C_{12}^*(n) \rightarrow A_{12}(n) \quad \text{et} \quad C_{21}^*(n) \rightarrow A_{21}(n)$$

2.4 Améliorations

En général, le gain d'adaptation peut varier au cours du temps et être différent pour les deux sorties. Il est souhaitable que la capacité de poursuite soit améliorée par le contrôle des gains d'adaptation. Une méthode simple est fondée sur l'algorithme LMS normalisé [6]. Dans ce cas, les deux gains d'adaptation sont normalisés par l'énergie du signal estimé à la sortie.

En principe, l'apprentissage de l'algorithme de séparation de sources se fait en permanence ce qui permet la poursuite des sources dans un mélange variable. Dans le cas de séparation de la parole et du bruit, une pause silencieuse existe souvent entre les syntagmes ou les phrases. L'énergie du signal primitif correspondant au silence est normalement très faible. En l'absence de signal la convergence de l'algorithme peut être perturbée, c'est pourquoi on a intérêt à supprimer l'adaptation pendant les silences.

3. Résultats de simulation

Des simulations de séparation de signal et de bruit ont été faites sur un mélange convolutif d'un signal provenant d'une base de données standardisée (BDSON) et d'un bruit pseudo-aléatoire.

Le mélange convolutif du signal échantillonné d'une phrase de parole et de bruit pseudo-aléatoire a été simulé à l'aide de deux filtres FIR passe-bas [5] avec des ordres de filtres de 10 à 20 et une fréquence d'échantillonnage Fe = 10 à 16 kHz. Les signaux de mélange sont calculés par les relations (1'). Initialement les coefficients $c_{ij}(k)$ des filtres sont mis à zéro. L'algorithme adaptatif peut séparer le signal et le bruit dans un mélange convolutif inconnu. La convergence de l'algorithme est obtenue au bout de 60 ms environ (600 pas de calcul). A la convergence, les deux filtres estimés sont voisins des deux filtres (inconnus de l'algorithme) du mélange. A la sortie du modèle, le signal de parole apparaît sur une sortie tandis que l'autre sortie fournit le bruit (Fig. 3). Le rapport signal/bruit (RSB) gagné en moyenne est de 19 dB et le RSB maximal à la sortie peut atteindre 28 dB (Fig. 4). Les améliorations de l'algorithme ont apporté de bons résultats en simulation, notamment dans le cas de mélanges non-stationnaires. Cependant, malgré ces résultats satisfaisants, on peut remarquer que l'erreur paramétrique sur les coefficients estimés des filtres reste élevée : de 10 à 20%. La figure 5 illustre l'erreur quadratique moyenne sur les deux filtres au cours de l'apprentissage, où chaque trame représente 200 échantillons.

4. Conclusion

Les résultats de notre simulation montrent qu'il est possible de

séparer des signaux à large bande dans un mélange convolutif inconnu par un algorithme de séparation aveugle de sources qui estime simultanément les filtres inverses. Le principe de l'algorithme est extrêmement intéressant puisqu'aucune hypothèse particulière sur des signaux n'est nécessaire et qu' on ne connaît ni le mélange ni les signaux primitifs. Dans tous les cas de simulation, la convergence de l'algorithme est atteinte après quelques trames de signal. Nous avons appliqué l'algorithme au cas réel dans lequel une source de bruit et un signal de parole sont enregistrés par deux microphones. Les résultats dans ce cas ne sont pas encore suffisants : nous considérons que le modèle simplifié de mélange convolutif n'est pas assez réaliste (le mélange inconnu est complexe : réverbération, écho, retard non entier, etc). Par ailleurs, l'hypothèse de sources ponctuelles est aussi susceptible d'être remise en cause. Il faut donc étudier maintenant l'algorihme dans un cas plus général de mélange convolutif modélisé par quatre filtres (Fig. 1).

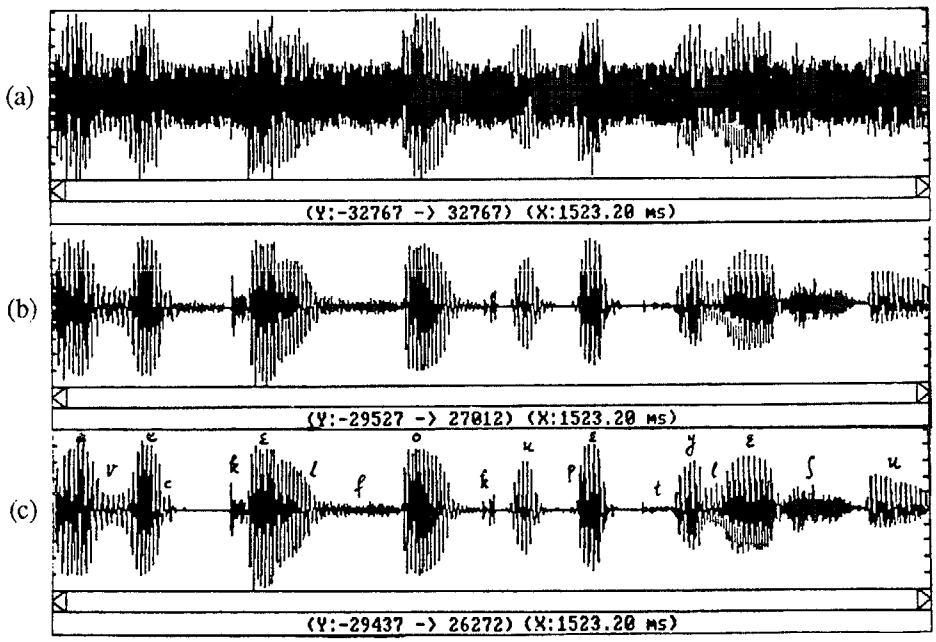

Fig. 3. Résultat expérimental :
(a) Signal de la parole bruité du mélange convolutif,
(b) Signal de la parole extrait à la sortie du modèle,
(c) Signal de la parole original : "Avec quelle faux coupais-tu les choux ?"

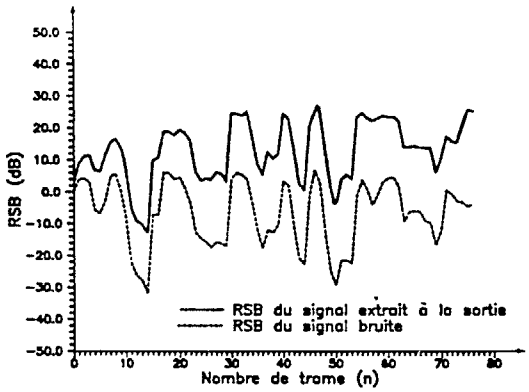

Fig. 4. RSB du signal bruité et du signal à la sortie.

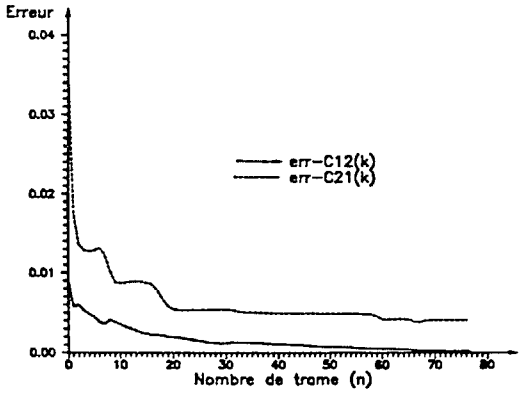

Fig. 5. Erreur quadratique moyenne des coefficients des filtres.

Références

[1] C. JUTTEN, "Calcul neuromimétique et traitement du signal, analyse en composantes indépendantes", Thèse d'état ès Sciences Physiques, USMG/ INPG, Grenoble 1987.
[2] B.WIDROW & S.D. STEARNS, Adaptive signal processing, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1985.
[3] M. FEDER, A. V. OPPENHEIM & E. WEINSTEIN, "Maximum likelihood noise cancellation using the EM algorithm", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-37, n° 2, February 1989.
[4] W. A. HARRISON, J. S. LIM & E. SINGER, "A new application of adaptive noise cancellation", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, n° 1, February 1986.
[5] P. M. PETERSON, "Simulating the response of multipe microphones to a single acoustic source in a reverberant room", J. Acoust. Soc. Am. 80 (5), Nov. 1986.
[6] O. MACCHI, A. GILLOIRE, C. SERVIERE et al., "Comparaison d'algorithmes adaptatifs en contexte non-stationnaire", Traitement du signal , Vol. 6, n°5, 1989.
[7] E. SOROUCHYARI, "Blind separation of sources. Part III: Stability Analysis". To appear in the signal processing.
[8] P.. COMMON, C. JUTTEN, J. HERAULT, "Blind separation of sources. Part II: Problem Statement". To appear in the signal processing.

Remerciements : Ce travail est partiellement financé dans le cardre du projet ESPRIT-BRA , n° 3049 : NERVES.