Séparation de Sources par lissage cepstral des masques binaires

Ibrahim Missaoui[1] Zied Lachiri[1, 2]
(1) École nationale d'ingénieurs de Tunis, ENIT, BP. 37 Le Belvedere, 1002 Tunis, Tunisie
(2) Institut national des sciences appliquées et de technologie,INSAT, BP 676 centre urbain cedex,Tunis, Tunisie
brahim.missaoui@enit.rnu.tn, zied.lachiri@enit.rnu.tn

RÉSUMÉ

Dans cet article, nous proposons un système de séparation des signaux de parole à partir de deux mélanges convolutifs. Le système suggéré est basé sur la combinaison d'une technique de séparation aveugle de sources avec une procédure de masquage temps-fréquence, suivie d'un lissage cepstral. En effet, après la séparation des signaux sources, les masques binaires estimés subissent un lissage cepstral afin de réduire les fluctuations des artefacts introduites par l'opération de masquage temps-fréquence. Les résultats d'évaluation ont montrés l'efficacité du système proposé même dans les cas les plus défavorables.

ABSTRACT

Source separation by cepstral smoothing of binary masks

In this paper, we propose a separation system of speech signals from two convolutive mixtures. The suggested system is based on the combination of blind source separation technique with a time-frequency masking procedure, followed by a smoothing cepstral. Indeed, after separation of signal sources, the estimated binary masks undergo a cepstral smoothing to reduce the fluctuations artifacts which introduced by time-frequency masking operation. The evaluation results have shown the effectiveness of the proposed system even in the most unfavorable case.

MOTS-CLÉS : Masque binaire idéal, Lissage cepstral, Séparation aveugle de sources.

KEYWORDS: Ideal binary mask, Cepstral smoothing, Blind source separation .

1 Introduction

Le problème de séparation aveugle de sources (SAS) consiste à extraire des signaux inconnus provenant de différentes sources, à partir de leurs mélanges, sans tenir compte d'aucune information à priori, ni sur la nature du mélange ni sur les signaux sources elles-mêmes.
Les approches de SAS développées pour traiter ce problème dans le cas convolutif peuvent classées en deux grandes catégories (Pedersen *et al.*, 2007) : ceux qui tendent de le résoudre dans le domaine temporel (Gorokhov et Loubaton, 1997; Douglas *et al.*, 2007) et ceux qui transforment ce problème dans le domaine fréquentiel (Parra et Spence, 2000; Makino *et al.*, 2005; Yoshioka *et al.*, 2009). Toutefois, parmi les algorithmes proposés dans la littérature, il

n'existe pas encore un algorithme fiable qui peut être utilisé pour les différents signaux mélanges, surtout dans le cas de réverbération et dans le cas bruité. La performance de séparation, dans ces deux cas, reste encore limitée et exige d'autre amélioration.
Dans ce sens, plusieurs méthodes de SAS basées sur le masquage temps-fréquence ont été développées (Yilmaz et Rickard, 2004; Sawada *et al.*, 2006). Ces méthodes consistent à appliquer un masque temps-fréquence binaire aux signaux mélanges. Récemment, la notion de masque binaire idéal a été introduite comme étant l'objectif principal de l'analyse de scènes auditives computationnelle (Wang et Brown, 2006). Cette technique a montré qu'il est bien adapté à la séparation de signaux de paroles. En fait, il a montré des propriétés remarquables dans la suppression d'interférences ainsi que dans l'amélioration de l'intelligibilité du signal cible (Wang *et al.*, 2009). Le masque binaire idéal est déterminé en comparant chaque unité temps-fréquence de signal cible avec celle d'interférence tout en associant une valeur 1 si l'énergie de cible est supérieure à celle d'énergie de l'interférence et une valeur 0 en cas inverse (Wang, 2005; Wang et Brown, 2006). Cependant, sans la connaissance à priori de signal de parole cible et celui d'interférence, l'estimation exacte d'un masque binaire idéal à partir de signaux mélanges devient une tâche difficile (Jan *et al.*, 2009; Madhu *et al.*, 2008).
Dans ce travail, nous proposons d'estimer les masques binaires à partir des signaux résultants d'une étape de séparation en utilisant un algorithme de SAS. Ces masques subissent ensuite une opération de lissage cepstral. Cette dernière permet de réduire les fluctuations des artefacts, connue sous le nom de "bruit musical", provoquées généralement par la masquage temps-fréquence (Jan *et al.*, 2009; Madhu *et al.*, 2008).
Ce papier est organisé comme suit : Nous commençons dans la section 2 par présenter le principe de SAS dans le cas convolutif. L'étape de lissage cepstral des masques binaires est détaillé dans la section 3. la section 4 expose les expériences et les mesures d'évaluations obtenues. Enfin, la section 5 conclure notre travail.

2 Séparation aveugle des signaux de parole

Dans le cas convolutif, le SAS consiste à extraire N signaux inconnues s_i, à partir de leurs mélanges x_j enregistrés par M microphones sans aucune information à priori. Le modèle mathématique associé à ce type des mélanges est définie comme suit :

$$x_j(m) = \sum_{i=1}^{N} \sum_{p=1}^{P} h_{ji}(p) s_i(m-p+1) \tag{1}$$

Avce h_{ji} sont les réponses impulsionnelles des filtres de mélange.
Ce modèle peut être écrite sous la forme matricielle suivante :

$$X(m) = H(m) * S(m) \tag{2}$$

Avec $X(m) = [x_1(m), .., x_M(m)]^T$ et $S(m) = [s_1(m), .., s_N(m)]^T$ sont définies comme étant le vecteur des signaux mélanges $x_j(m)$ et celui des signaux sources $s_i(m)$, $*$ est l'opérateur de convolution et $H(m)$ est la matrice des filtres de mélange.
En appliquant la transformée de Fourier à court terme à l'équation (1), le problème de SAS convolutif est transformé en un ensemble des problèmes instantanés dans le domaine fréquentiel (Parra et Spence, 2000; Makino *et al.*, 2005; Yoshioka *et al.*, 2009). Ce qui donne l'équation

suivante :

$$X(k,m) = H(k)S(k,m) \tag{3}$$

l'objectif de SAS consiste à trouver une matrice des filtres $W(k)$ qui sera ensuite utilisé pour extraire les signaux sources à partir des mélanges comme suit :

$$\hat{S}(k,m) = W(k)X(k,m) \tag{4}$$

Les signaux séparés $\hat{S}(m) = [\hat{s}_1(m),..,\hat{s}_N(m)]^T$ sont obtenus en appliquant la transformée de Fourier à court terme inverse à la représentation temps-fréquence des ces signaux $\hat{S}(k,m) = [\hat{s}_1(k,m),..,\hat{s}_N(k,m)]^T$. Dans ce travail, nous traitons le cas de deux mélanges convolutifs où chaque mélange est formé par deux signaux de parole ($N = M = 2$).
Le système de séparation proposé, présenté par la figure 1, comporte deux modules. Dans le premier module, les signaux séparés sont extraits à l'aide de l'algorithme de SAS développé par Parra et Spence (Parra et Spence, 2000). Cet algorithme est basé sur l'exploitation de la non stationnarité de signal de parole. Il permet de déterminer la matrice de filtres $W(k)$ en effectuant une diagonalisation simultanée du spectre de puissance croisée. Cette matrice des filtres est ensuite utilisée pour obtenir les signaux séparés. Le deuxième module correspond à l'étape de lissage cepstral des masques binaires. Ce module comporte deux étapes. Dans la première étape, deux masques binaires sont estimés à partir des signaux séparés obtenus dans le module précédant. Ensuite, une étape de lissage temporel de ces deux masques est réalisée dans le domaine cepstral afin de réduire les fluctuations des artefacts introduites par l'opération de masquage temps-fréquence. Les deux masques lissés sont ensuite converti en domaine spectral et appliqués aux deux signaux dans le but d'obtenir une estimation finale de signaux sources. Nous décrivons dans le paragraphe suivant l'étape de lissage cepstral des masques binaires.

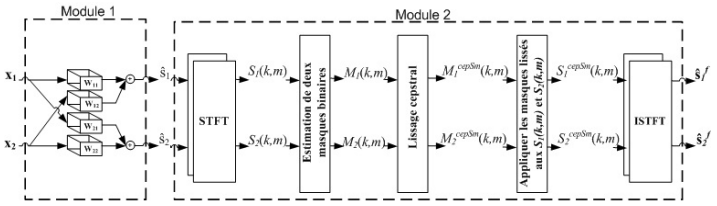

FIGURE 1 – Le système de séparation proposé

2.1 Les masques binaires

Les signaux séparés $\hat{s}_1$ et $\hat{s}_2$ obtenues dans le premier module sont transformés dans le domaine temps-fréquence en utilisant la transformée de Fourier à court terme. Les deux spectrogrammes correspondants sont notés par $S_1(k,m)$ et $S_2(k,m)$.

$$\begin{aligned} \hat{s}_1 &\rightarrow S_1(k,m) \\ \hat{s}_2 &\rightarrow S_2(k,m) \end{aligned} \tag{5}$$

Les deux masques binaires idéals M_1 et M_2 sont estimés en comparant l'énergie de chaque zone temps-fréquence de ces deux spectrogrammes comme suit :

$$\begin{array}{l} M_1(k,m) = \left\{ \begin{array}{ll} 1 & \text{Si } \left|S_1(k,m)\right| \succ \left|S_2(k,m)\right| \\ 0 & \text{Sinon} \end{array} \right. \\ M_2(k,m) = \left\{ \begin{array}{ll} 1 & \text{Si } \left|S_2(k,m)\right| \succ \left|S_1(k,m)\right| \\ 0 & \text{Sinon} \end{array} \right. \end{array} \tag{6}$$

2.2 Lissage cepstral des masques binaires

Afin de réduire les artéfacts musicaux produits généralement par la technique de masquage temps-fréquence, les deux masques binaires sont transformés en domaine cepstral dans lequel plusieurs niveaux de lissage temporel sont effectués (Oppenheim et Schafer, 2009). Cette procédure de lissage cepstral qui se base sur le mécanisme de production de parole, permet de réduire le bruit musical, tout en préservant la structure à large bande et l'information harmonique du signal de parole cible (Jan *et al.*, 2009; Madhu *et al.*, 2008; Oppenheim et Schafer, 2009). La représentation cepstral de chacun de deux masques spectraux M_1 et M_2 est donnée par l'équation suivante :

$$M_i^{cep}(l,m) = DFT^{-1}\{\ln(M_i(k,m))\mid_{k=1,..,K-1}\}, \quad i = 1,2 \tag{7}$$

Avec l est l'indice des bins fréquentiels et K est la longueur de la transformée de Fourier discrète (TFD) (Jan *et al.*, 2009; Madhu *et al.*, 2008). En appliquant un lissage temporel récursif du premier ordre aux masques résultants, Les deux masques lissés $\bar{M}_i^{cep}(l,m)$ sont donées par :

$$\bar{M}_i^{cep}(l,m) = \beta_l \bar{M}_i^{cep}(l,m-1) + (1-\beta_l) M_i^{cep}(l,m) \tag{8}$$

Avec la valeur de paramètre de niveau de lissage β_l est choisie en fonction des valeurs de l'indice des bins fréquentiels l comme suit :

$$\beta_l = \left\{ \begin{array}{l} \beta_{env} \quad \text{if } l \in \{0,..,l_{env}\} \\ \beta_{pitch} \quad \text{if } l = l_{pitch} \\ \beta_{peak} \quad \text{if } l \in \left\{\left(l_{env}+1\right),...,K\right\} \backslash l_{pitch} \end{array} \right. \tag{9}$$

Où $0 < \beta_{env} < \beta_{pitch} < \beta_{peak} < 1$ et le symbole "\" désigne l'exclusion de l_{pitch} de l'intervalle $[l_{env}+1;K]$.
Pour de petites valeurs de l'indice des bins fréquentiels, les valeurs correspondants de $\bar{M}_i^{cep}(l,m)$ représentent l'enveloppe spectral du masque $M_i(k,m)$ (Madhu *et al.*, 2008; Oppenheim et Schafer, 2009). Pour cela, le paramètre β_{env} est fixé à une petite valeur afin d'éviter la distorsion de l'enveloppe spectrale. De même, la structure harmonique du signal est maintenue en appliquant un faible lissage β_{pitch} pour $l = l_{pitch}$. Le reste des valeurs de l'indice des bins fréquentiels contient les pics spectraux aléatoires indésirables (Oppenheim et Schafer, 2009). Ces pics engendrent généralement la distorsion harmonique. Par conséquent, un fort lissage (β_{peak})) dans cette région est exigé afin de réduire les artefacts (Madhu *et al.*, 2008; Oppenheim et Schafer, 2009).
La fréquence fondamentale l_{pitch} est calculée pour chaque fenêtre temporelle m à partir de signaux séparés $\hat{s}_1$ et $\hat{s}_2$ comme suit (Jan *et al.*, 2009) :

$$l_{pitch} = \arg\max_l \left\{ sig^{cep}(l,m) | l_{low} \leq l \leq l_{high} \right\} \tag{10}$$

Avec $sig^{cep}(l,m)$ est la représentation cepstrale de signal séparé obtenue par le module 1. Les deux valeurs de l_{low} et l_{high} sont choisies de sorte que l'intervalle correspondant puisse accueillir les fréquences fondamentales de la voix humaine entre 50 to 500 Hz.
La version lissée du masque spectrale est calculée selon l'équation suivante :

$$M_i^{cepSm}(k,m) = \exp\left(DFT\left\{\bar{M}_i^{cep}(l,m)|_{l=0,\ldots,K-1}\right\}\right) \tag{11}$$

Le masque lissés est ensuite appliqués à la représentation temps-fréquence $S_i(k,m)$ de signal séparé obtenue par le module 1.

$$S_i^{cepSm}(k,m) = M_i^{cepSm} S_i(k,m) \tag{12}$$

Enfin, les signaux estimés finales sont récupérés dans le domaine temporel en utilisant la transformée de Fourier à court terme inverse.

3 Résultats expérimentaux

Pour évaluer la performance du système proposé, nous avons utilisé plusieurs configurations de mélanges convolutifs artificiellement établis, où chaque mélange est formé par deux signaux de parole. Dans ce papier, nous présentons les résultats obtenues par deux expériences. Dans la première expérience, les deux signaux mélanges sont formés en utilisant des canaux convolutifs, alors que dans la deuxième expérience, nous mélangeons deux signaux de parole à l'aide d'une simulation d'une salle acoustique établie par Allen et Berklen (Gaubitch, 1979). Les valeurs des différents paramètres de notre système de séparation est présentés dans le tableau 1.

DFT length= 2048	$\beta_{env} = 0$	$l_{env} = 8$
overlap factor=0.75	$\beta_{pitch} = 0.9$	$l_{low} = 16$
	$\beta_{peak} = 0.4$	$l_{high} = 120$

TABLE 1 – Les valeurs des paramètres utilisées

L'évaluation de notre système de séparation porte sur la qualité de séparation à travers un critère de performance fournie par la boite à outils d'évaluation "BSS EVAL toolbox", en particulier le rapport le rapport signal à interférence (SIR) (Vincent *et al.*, 2006). En outre, la qualité de signaux séparés est évaluée en utilisant l'indice de qualité PESQ (Perceptual Evaluation of Speech Quality). Ce dernier représente l'équivalence de mesure subjective de Mean Opinion Score (MOS) (ITU-TP.862, 2001), . Les résultats des évaluations obtenues sont comparés à ceux obtenus par algorithme de Parra (Parra et Spence, 2000).

– **Expérience 1** :Dans la première expérience, les signaux mélanges sont obtenus en appliquant, aux deux signaux de parole, quatre canaux convolutifs définies par l'équation (13) . Les signaux utilisé sont issues de base TIMIT (Fisher *et al.*, 1986).

$$\begin{aligned} h_{11}(m) &= [1.0, 0.8, 0.7, 0.4, 0.3, 0.25, 0.2, 0.15] \\ h_{12}(m) &= [0.6, 0.5, 0.5, 0.4, 0.3, 0.2, 0.25, 0.1] \\ h_{21}(m) &= [0.5, 0.5, 0.4, 0.35, 0.3, 0.3, 0.2, 0.1] \\ h_{22}(m) &= [1.0, 0.9, 0.8, 0.6, 0.4, 0.35, 0.3, 0.15] \end{aligned} \tag{13}$$

Les canaux de mélange sont choisis les mêmes que celle utilisée dans (Rahbar et Reilly, 2001) et (Mei *et al.*, 2008).

	SIR		PESQ	
	Algorithme de Parra	SP	Algorithme de Parra	SP
signal 1	20.71 dB	25.74 dB	2.92	3.06
signal 2	14.92 dB	18.05 dB	3.13	3.33
Moyenne	17.81 dB	21.98 dB	3.02	3.19

TABLE 2 – Les valeurs de SIR et PESQ obtenues en utilisant le système proposé (SP) et l'algorithme de Parra

Le tableau 2 présente les résultats de rapport SIR et l'indice de qualité PESQ obtenus, dans la première expérience, en utilisant le système proposé et l'algorithme de Parra. Nous remarquons que notre système fournie un bon résultat par rapport à celui de l'algorithme de Parra pour les deux signaux. En effet, nous avons enregistré une valeur moyenne de SIR d'ordre de 21.98 dB en utilisant notre système de séparation et 17.81 dB en utilisant l'algorithme de Parra. Nos résultats sont confirmés par l'amélioration de l'indice de qualité PESQ. Nous avons obtenus une valeur moyenne de PESQ égale 3,19 pour notre système et 3,02 pour l'algorithme de Parra.

– **Expérience 2 :** Dans la deuxième expérience, notre système est testé sur des mélanges convolutifs fournis à l'aide d'une simulation d'une salle acoustique réverbérant établie par Allen et Berklen (Gaubitch, 1979). Chaque mélange est formé par deux signaux de parole mélangés pour différents valeurs de temps de réverbération RT (RT=30,50,100,150,200 ms). Les signaux de parole utilisés, ayant approximativement le même niveau d'intensité sonore et un longueur de 5 secondes, sont échantillonné à 10 KHz (Pedersen *et al.*, 2008).

		SIR(dB)		PESQ	
RT (ms)		Algorithme de Parra	SP	Algorithme de Parra	SP
30	signal 1	20.75	26.68	2.83	2.93
	signal 2	20.99	36.13	3.27	3.42
	Moyenne	20.87	31.04	3.05	3.67
50	signal 1	21.08	26.88	2.57	2.62
	signal 2	17.93	29.15	3.22	3.34
	Moyenne	19.50	28.01	2.89	2.98
100	signal 1	12.66	20.78	1.94	1.94
	signal 2	17.61	27.54	2.79	2.90
	Moyenne	15.13	24.16	2.36	2.42
150	signal 1	13.83	29.10	1.71	1.68
	signal 2	2.33	8.64	2.50	2.65
	Moyenne	8.02	18.87	2.10	2.16
200	signal 1	3.72	17.29	1.60	1.66
	signal 2	-0.72	7.51	2.36	2.42
	Moyenne	1.5	12.4	1.98	2.04

TABLE 3 – Les valeurs de SIR et PESQ obtenues en utilisant le système proposé (SP) et l'algorithme de Parra pour différents valeurs de RT.

Les résultats d'évaluation de cette série des tests obtenus en utilisant le système proposé et

l'algorithme de Parra, sont récapitulé dans le tableau 3. Nous constatons que notre système fournie un bon résultat en terme de SIR, pour les différentes valeurs de RT, par rapport à ceux obtenus par l'algorithme Parra. Par exemple, la valeur moyenne de SIR pour RT=30 est de 20,87 dB en utilisant l'algorithme de Parra alors que notre système fournie un rapport SIR égale à 31,04 dB. Cette amélioration est confirmée par la mesure de l'indice de qualité PESQ qui permet d'évaluer la qualité des signaux séparés. Nous remarquons que notre système a fournie des résultats remarquables en termes de PESQ. Par exemple, pour RT=30 ms, nous avons obtenue une valeur de PESQ égale à 2.93 tandis que l'algorithme de Parra fournie une valeur de l'ordre de 2.83.
D'après le tableau 3, la meilleure performance de système suggéré est obtenue pour les petites valeurs de RT. Cette performance se dégrade progressivement en augmentant la valeur de RT de 30 à 200 ms. Ce résultat est dû à l'augmentation des réflexions sonores pour les hautes valeurs de RT.

4 Conclusion

Nous avons proposé un système de séparation basé sur la technique de séparation aveugle de sources et la procédure de masquage temps-fréquence, suivie d'une opération de lissage cepstral. Les signaux séparés obtenus en utilisant un algorithme de SAS, sont exploités pour estimer deux masques binaires. Ces masques ont subies ensuite un lissage cepstral afin de réduire les fluctuations des artefacts introduits par l'opération de masquage temps-fréquence. Les résultats de séparation obtenus sont très encourageants et montrent une considérable amélioration de la qualité des signaux séparés ainsi que la réduction des fluctuations des artefacts.

Références

DOUGLAS, S., GUPTA, M., SAWADA, H. et MAKINO, S. (2007). Spatio-temporal fastica algorithms for the blind separation of convolutive mixtures. *IEEE Transactions on Audio Speech Lang. Processing*, 15(5):1511–1520.

FISHER, W., DODINGTON, G. et GOUDIE-MARSHALL, K. (1986). The timit-darpa speech recognition research database : Specification and status. *In DARPA Workshop on Speech Recognition*.

GAUBITCH, N. (1979). Allen and berkeley image model for room impulse response. *In Imperial College London*.

GOROKHOV, A. et LOUBATON, P. (1997). Subspace based techniques for second order blind separation of convolutive mixtures with temporally correlated sources. *IEEE Transactions on Circuit Systems I : Fundamental Theory and Applications*, 44(9):813–820.

ITU-TP.862 (2001). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. International Telecommunication Union, Geneva.

JAN, T., WANG, W. et WANG, D. (2009). A multistage approach for blind separation of convolutive speech mixtures. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1713–1716.

MADHU, N., BREITHAUPT, C. et MARTIN, R. (2008). Temporal smoothing of spectral masks in the cepstral domain for speech separation. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 45–48.

MAKINO, S., SAWADA, H., MUKAI, R. et ARAKI, S. (2005). Blind source separation of convolutive mixtures of speech in frequency domain. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E88-A*, 7:1640–1655.

MEI, T., MERTINS, A., YIN, F., XI, J. et CHICHARO, J. (2008). Blind source separation for convolutive mixtures based on the joint diagonalization of power spectral density matrices. *Signal Processing*, 88(8):1990–2007.

OPPENHEIM, A. et SCHAFER, R. (2009). *Discrete Time Signal Processing*. Prentice Hall, New Jersey, third édition.

PARRA, L. et SPENCE, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327.

PEDERSEN, M., LARSEN, J., KJEMS, U. et PARRA, L. C. (2007). A survey of convolutive blind source separation methods. *In Springer Handbook of Speech Processing*, pages 1–34. Springer Press.

PEDERSEN, M., WANG, D., LARSEN, J. et KJEMS, U. (2008). Two-microphone separation of speech mixtures. *IEEE Transactions on Neural Networks*, 19:475–492.

RAHBAR, K. et REILLY, J. (2001). Blind source separation of convolved sources by joint approximate diagonalization of crossspectral density matrices. *In IEEE International Conference on Acoustics Speech and Signal Processing*, pages 2745–2748.

SAWADA, H., ARAKI, S., MUKAI, R. et MAKINO, S. (2006). Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Trans. Audio, Speech, Lang. Process*, 14(6): 2165–2173.

VINCENT, E., GRIBONVAL, R. et FEVOTTE, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469.

WANG, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. *In* DIVENYI, P, éditeur : *Speech Separation by Humans and Machines*, pages 181–197. Springer.

WANG, D. et BROWN, G. (2006). *Computational Auditory Scene Analysis : Principles, Algorithms, and Applications*. Wiley-IEEE Press, New Jersey.

WANG, D., KJEMS, U., PEDERSEN, M., BOLDT, J. et LUNNER, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *Journal of the Acoustical Society of America*, 125:2336–2347.

YILMAZ, O. et RICKARD, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847.

YOSHIOKA, T., NAKATANI, T. et MIYOSHI, M. (2009). Fast algorithm for conditional separation and dereverberation. *In Proc 17th European Signal Processing Conference*, pages 1432–1436.