Amélioration de l'intelligibilité de la parole dans des enregistrements de « boîtes noires aéronautiques » à l'aide de méthodes de Séparation Aveugle de Sources

Benjamin Bigot[1] Hélène Devulder[1, 2] Matthieu Puigt[2]
(1) Bureau d'Enquêtes et d'Analyses (BEA), 10 rue de Paris, 93352 Le Bourget, France
(2) Univ. Littoral Côte d'Opale, LISIC – UR 4491, 50 rue F. Buisson, 62228 Calais, France
benjamin.bigot@bea.aero, com@bea.aero, matthieu.puigt@univ-littoral.fr

RÉSUMÉ
Le BEA est l'autorité française en charge des enquêtes de sécurité en cas d'incidents ou d'accidents d'aéronefs civils, et dans ce contexte procède à l'analyse des fichiers audio des enregistreurs de conversations (CVR). Les contraintes de conception des CVR ont pour conséquence la présence d'une quantité importante de zones de parole superposée, ce qui complique l'exploitation de ces données. Dans cet article nous présentons une première étude sur l'apport potentiel de la Séparation Aveugle de Sources pour l'amélioration de l'intelligibilité des enregistrements CVR.

ABSTRACT
Improving speech intelligibility of Cockpit Voice Recorders with speech source separation.

The BEA is the French authority in charge of safety investigations in the event of a civil aircraft accident, and in this context analyses audio files from cockpit voice recorders (CVRs). The CVR design constraints result in the presence of a large amount of superimposed speech areas, which complicates the use of this data. In this article we present a first study on the potential contribution of Blind Source Separation for improving the intelligibility of CVR recordings.

MOTS-CLÉS : séparation de sources de parole ; boites noires aéronautiques.

KEYWORDS: speech source separation ; cockpit voice recorder.

1 Introduction

Le Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA) est l'autorité française en charge des enquêtes de sécurité lors d'incidents ou d'accidents d'aéronefs civils. Le BEA réalise l'exploitation des enregistreurs de vol – plus communément appelés « boîtes noires » – qui comportent un enregistreur de conversations (CVR pour *Cockpit Voice Recorder*) et un enregistreur de paramètres (FDR pour *Flight Data Recorder*). L'analyse et la transcription des CVR sont réalisées au bénéfice exclusif de l'enquête de sécurité par des enquêteurs spécialisés.

Les causes de dégradation de l'intelligibilité de la parole dans les enregistrements CVR sont multiples. En premier lieu, la conception même des enregistreurs de conversations amène à trouver, sur les 4 canaux audio enregistrés simultanément, une forte proportion de parole superposée. La parole superposée est également potentiellement plus fréquente dans des enregistrements d'accidents ou d'incidents durant lesquels les activités vocales et sonores peuvent se densifier. Ces phénomènes

10.21437/JEP.2022-35

compliquent le travail des analystes audio et peuvent dans le pire cas mener à une perte d'informations cruciales pour l'enquête de sécurité. Le BEA utilise déjà des algorithmes spécifiques de soustraction de sources sonores et souhaite étudier l'apport de la Séparation Aveugle de Sources (SAS) pour l'amélioration de l'intelligibilité de la parole dans les enregistrements CVR.

La SAS est un problème générique dont les premiers travaux ont été proposés en France au milieu des années 1980 (Comon & Jutten, 2010). Une application classique de la SAS en audio est le problème du *« Cocktail Party »* qui consiste à séparer N sources (locuteurs) inconnues à partir des observations captées par M microphones distants et contenant des mélanges des sources. Ce travail est à notre connaissance la première application des méthodes de SAS sur des enregistrements réels de CVR. L'évolution des tâches proposées lors de challenges tels que CHiME (Watanabe *et al.*, 2020) semble indiquer un intérêt manifeste de la communauté de recherche pour des applications réalistes et ambitieuses, et motive le BEA à communiquer sur ses travaux de recherche et développement.

Après un aperçu de la SAS en section 2, nous présentons en section 3 le système audio CVR et ses caractéristiques ce qui nous mène à proposer dans cette même section un modèle des mélanges de sources sonores et à identifier un sous-ensemble de méthodes de SAS. L'apport de la SAS pour la tâche de transcription, évaluée par les enquêteurs spécialisés du BEA sur des enregistrements réels, est présentée en section 3.3. Nous concluons avec les perspectives de ces travaux en section 4

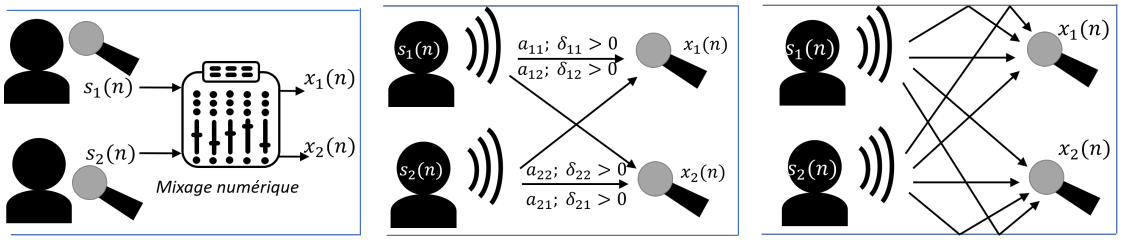
2 Aperçu de la Séparation Aveugle de Sources

Le problème fondamental de la SAS s'écrit formellement par l'Eq. (1), avec $\mathbf{X}$ la matrice d'observations disponible de taille $M \times T$ (où T est le nombre d'échantillons temporels), $\mathcal{A}$ l'opérateur de mélange et $\mathbf{S}$ la matrice de sources de taille $N \times T$ à estimer. Le canal de propagation de chaque source vers chaque capteur est supposé inconnu. Dans la pratique, le modèle de mélange de sources et le nombre de sources du mélange sont souvent connus, et il est fréquent de formuler des hypothèses sur des propriétés des signaux sources avant de réaliser leur estimation à partir des observations.

$$\mathbf{X} = \mathcal{A}(\mathbf{S}). \tag{1}$$

Les modèles de mélange considérés classiquement par les méthodes de SAS sont représentés en Fig. 1. Le mélange linéaire instantané (LI) en Fig. 1a considère uniquement les trajets directs sans délai de propagation entre les sources et les microphones. Le modèle de mélange anéchoïque en Fig. 1b prend en compte les différences de délais de propagation entre les sources et les capteurs. Le modèle de mélange convolutif en Fig. 1c considère les multiples trajets d'une source vers les microphones.

La littérature scientifique rassemble plusieurs **familles de méthodes de SAS** en fonction du type de modèle de mélange considéré. L'hypothèse historique de la SAS est l'indépendance statistique des sources sur laquelle repose l'*analyse en composantes indépendantes* (ICA pour Independent Component Analysis en anglais) (Comon & Jutten, 2010). Moins adaptée aux mélanges de parole que d'autres familles de méthodes (Abrard & Deville, 2005; Puigt *et al.*, 2009), l'ICA n'est pas retenue dans notre étude. Les méthodes d'*analyse en composantes parcimonieuses* (SCA pour Sparse component Analysis en anglais) reposent sur l'hypothèse de parcimonie des sources. Un signal de variance non nulle est parcimonieux si une proportion non-négligeable de ses échantillons est nulle ou quasi-nulle dans l'espace de représentation des paramètres (Puigt, 2007). Cette hypothèse, pertinente pour les signaux de parole et de musique, a mené à de nombreuses contributions (Yilmaz & Rickard,



(a) mélange linéaire instantané (b) mélange anéchoïque (c) mélange convolutif

FIGURE 1 – Modèles de mélange classiques de la SAS.

2004; Abrard & Deville, 2005). Des méthodes de *SAS fondée sur la factorisation en matrices non-négatives* (NMF-SAS) se sont imposées à partir du milieu des années 2000. La NMF-SAS propose de résoudre un problème linéaire de réduction de dimension spécifique (Lee & Seung, 1999; Smaragdis & Brown, 2003; Ozerov & Févotte, 2010) en modélisant une matrice de coefficients non-négatifs par un produit de matrices non-négatives. La NMF-SAS a à plusieurs reprises été combinée à des méthodes d'apprentissage (Wilson *et al.*, 2008). L'état de l'art s'articule désormais quasiment exclusivement autour d'*approches reposant sur l'apprentissage profond* (Nugraha *et al.*, 2016).

L'estimation des sources par des méthodes de SAS ne peut être réalisée qu'à un facteur d'échelle ou de filtre près, et à une permutation près. Nous considérons par la suite les données dans un espace de représentation temps-fréquence selon une décomposition atomique obtenue par transformée de Fourier à court terme (TFCT). La relation (1) s'écrit alors $\mathbf{X}(\omega, n) = \mathcal{A}\left(\mathbf{S}(\omega, n)\right)$, où ω et n représentent respectivement la pulsation et l'indice temporel considérés dans le domaine temps-fréquence. Selon les méthodes de SAS considérées, la reconstruction est basée sur des principes différents. La plupart des approches de SCA suivent une structure en deux étapes dans lesquelles (i) l'opérateur de mélange est estimé, puis (ii) les sources sont estimées sous forme d'un problème inverse. Dans ce cadre, l'*inversion des filtres de mélange estimés* est la méthode de reconstruction la plus simple et peut être utilisée lorsque le nombre de sources est inférieur ou égal au nombre d'observations. Lorsque les mélanges considérés sont convolutifs, une stratégie classique consiste à supposer un mélange LI à valeurs complexes dans chaque bande fréquentielle, c.-à-d. $\mathbf{X}(\omega, n) = \mathbf{A}(\omega)\mathbf{S}(\omega, n)$. Après la première étape d'une approche de SCA où, pour chaque pulsation ω, la matrice $\mathbf{A}(\omega)$ est connue et supposée inversible, il devient possible de déduire $\mathbf{S}(\omega, n)$ à partir de $\mathbf{X}(\omega, n)$ et de $\mathbf{A}(\omega)^{-1}$. Cependant, le filtre estimé de $\mathbf{A}(\omega)$ est valide à une indétermination de filtre près et à une permutation près pour chaque bande de fréquence. L'ordre dans lequel les sources sont estimées diffère donc d'une bande de fréquence à l'autre et il est nécessaire de réaliser des permutations sur $\mathbf{A}(\omega)$ préalablement à la reconstruction pour préserver l'ordre d'apparition des signaux dans les sources reconstruites. La *reconstruction par masquage binaire dans le domaine spectral* (Yilmaz & Rickard, 2004) suppose que dans chaque atome (ω, n) de l'espace de représentation des observations, une seule source est présente (ou dominante). Cette seconde méthode propose la construction de N masques binaires de dimensions similaires à celles des observations dans le domaine temps-fréquence utilisés dans le but de distribuer les atomes temps-fréquence en N représentations qui par application de la TFCT inverse mèneront à la production des N sources estimées. Dans la pratique le masquage binaire peut se révéler performant en dehors de cette hypothèse stricte en compensant les imprécisions des coefficients de la matrice de mélange estimée $\mathbf{A}(\omega)$. Par contre cette méthode présente alors l'inconvénient d'introduire un fort bruit musical dans les sources reconstruites. Ce bruit musical augmente avec le nombre de sources mélangées. D'autres stratégies ont été proposées pour la NMF-SAS qui n'utilise

que les modules des TFCT. Chaque observation y est décomposée comme une somme de signatures spectrales caractéristiques des sources (Smaragdis & Brown, 2003). Quand plusieurs observations sont disponibles, les filtres de mélange $A(\omega)$ sont aussi estimés et la NMF-SAS estime les modules de chaque source. Les phases des sources sont ensuite estimées par *filtrage de Wiener* des observations (Ozerov & Févotte, 2010).

3 Application de la SAS aux mélanges CVR

3.1 Le système audio CVR

Un CVR, dont une illustration est fournie Fig. 2a, est un dispositif enregistrant simultanément 4 canaux audio. Auparavant enregistrées sur bande magnétique, les données sont depuis les années 1990 numérisées et stockées sur des cartes mémoires placées dans un boîtier renforcé. Depuis le 1^{er} janvier 2021, les CVR équipant les nouvelles générations d'avion doivent protéger au minimum 25 heures d'enregistrement audio. La réglementation définit le contenu des canaux enregistrés par les CVR :

- canal n°1 : les signaux émis et reçus par le système audio du pilote en place gauche ;
- canal n°2 : les signaux émis et reçus par le système audio du pilote en place droite ;
- canal n°3 : les signaux émis et reçus par le système audio du poste $3^{\text{ème}}$ homme ;
- canal n°4 : le microphone d'ambiance en Fig. 2b (CAM pour *Cockpit Area Microphone*).

Avant d'être enregistrés sur la carte mémoire de l'enregistreur, les signaux provenant des voies pilotes et du CAM sont sous-échantillonnés respectivement à 7 kHz et 12 kHz, puis compressés au standard ADPCM (Adaptive Differential Pulse Code Modulation).

(a) un type de CVR

(b) un type de CAM

(c) un type d'équipement de tête

FIGURE 2 – Exemples de composants du système audio CVR

Le CAM est un microphone omnidirectionnel installé généralement sur le plafond du cockpit entre les pilotes (voir Fig. 2b). Il est connecté au CVR sans passer par le système audio de l'avion. Le CAM permet la captation des conversations et de l'ambiance sonore du cockpit, dont les signatures spectrales des groupes moto-propulseurs. Ce canal audio n'est pas considéré dans cette étude.

Les trois autres canaux contiennent des combinaisons des signaux reçus et émis à chaque poste du cockpit, ce qui correspond concrètement à une superposition des sons entendus dans le casque et ceux captés par les microphones de chaque membre d'équipage. Pour chaque canal, ces signaux sont mélangés par le système audio de l'avion, qui adapte dynamiquement les niveaux relatifs du casque et du microphone afin de garantir l'intelligibilité de la voix du pilote dans le mélange enregistré.

Les sources sonores disponibles dans les casques des pilotes sont typiquement les sons perçus par les microphones des autres postes d'équipage, les messages reçus du contrôle aérien et des autres aéronefs sur la fréquence radio, et les échanges avec les personnels navigants commerciaux. Les pilotes n'entendent pas leur propre voix dans leur casque, sauf lorsqu'ils émettent sur le canal radio. L'activation ainsi que le niveau de ces sources sonores dans un casque sont ajustés par chaque pilote grâce à un panneau de réglage individuel. Toutefois ces réglages ne sont pas perceptibles dans le CVR pour lequel les niveaux sonores de chaque source ont été figés lors de l'installation de l'enregistreur sur l'avion. Les niveaux relatifs des sources entendus dans le CVR ne reflètent donc pas les réglages des pilotes. Les signaux envoyés dans le casque d'un pilote en place gauche ou droite sont également reproduits par des haut-parleurs situés respectivement à l'avant gauche ou droit du cockpit.

Chaque poste pilote est équipé d'un casque-micro comme illustré Fig. 2c, d'un microphone à main et d'un microphone monté à l'intérieur d'un masque à oxygène. Ces microphones captent principalement la voix du pilote qui les porte, mais il est courant d'y percevoir à un niveau moindre l'environnement audio du cockpit et notamment les alertes sonores émises par les haut-parleurs du cockpit. Dans certaines configurations, les pilotes doivent activer leur microphone par une action sur un alternat. À la demande des enquêteurs de sécurité, une fonction « microphone ouvert » permet au CVR d'enregistrer les sons captés par les microphones indépendamment de l'état ouvert ou fermé de l'alternat.

3.2 Modèles de mélange CVR

Les analystes audio du BEA sont confrontés à des segments audio sur lesquels se superposent jusqu'à quatre sources sonores. Dans cette étude, nous nous limiterons aux mélanges de deux sources sonores, en considérant d'une part les superpositions des voix des deux pilotes en poste, et d'autre part le cas où la voix d'un pilote est couverte par un message reçu sur la fréquence radio. L'installation de *microphones ouverts* dans les avions de transport public rendent ces deux scénarios très fréquents.

Nous représentons en Fig. 3, le système complet correspondant à ces deux scénarios. Les signaux $S_1(\omega, n)$ et $S_2(\omega, n)$ correspondent respectivement aux signaux de parole des pilotes en place gauche et en place droite. Les signaux $HP_1(\omega, n)$ et $HP_2(\omega, n)$ sont les signaux émis par les haut-parleurs situés généralement à l'avant gauche et droit du cockpit. Dans les deux scénarios considérés, un premier mélange acoustique des sources a lieu dans le cockpit de l'avion. Ce mélange est assignable au modèle de mélange convolutif présenté précédemment. Dans un second temps, pour chaque poste de pilotage, les signaux provenant du microphone et du casque sont mélangés par le système audio

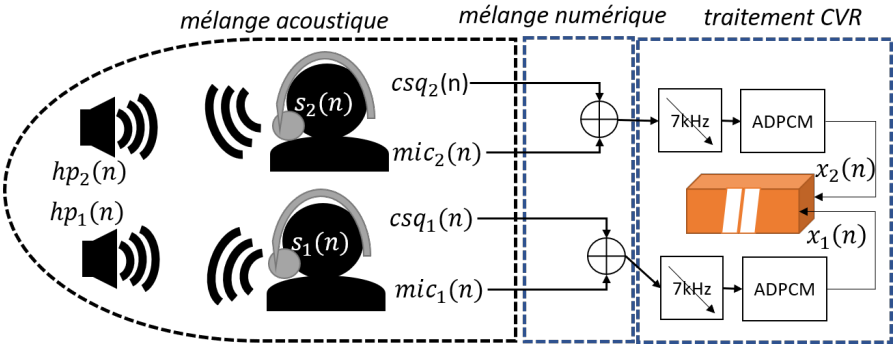


FIGURE 3 – Le mélange des sources dans le CVR

de l'avion. Cette superposition de signaux correspond à un modèle de mélange linéaire instantané. Les signaux combinés sont enfin envoyés respectivement vers les canaux n°1 et n°2 du CVR qui correspondent aux observations $X_1(\omega, n)$ et $X_2(\omega, n)$ de l'Eq. (1) de la SAS. Dans le cas du scénario considérant la parole superposée des pilotes, l'observation $X_i(\omega, n)$ – avec $i = 1, 2$ – peut être exprimée comme la somme des sons captés par le microphone et retranscris dans le casque i :

$$X_i(\omega, n) = MIC_i(\omega, n) + CSQ_i(\omega, n) = MIC_i(\omega, n) + \lambda_i MIC_{3-i}(\omega, n), \quad (2)$$

où λ_i est un coefficient d'atténuation introduit lors du mélange numérique réalisé par le système audio de l'avion sur le canal i (*cf.* Fig. 3). En considérant ces mélanges dans le domaine temps-fréquence, la superposition de sources captées par chaque microphone peut s'écrire

$$MIC_i(\omega, n) \approx A_{i1}(\omega)S_1(\omega, n) + A_{i2}(\omega)S_2(\omega, n) + A_{i3}(\omega)HP_1(\omega, n) + A_{i4}(\omega)HP_2(\omega, n), \quad (3)$$

où $A_{ij}(\omega)$ défini pour $j = 1, \ldots, 4$ correspond à la transformée de Fourier d'un filtre de propagation d'une source sonore vers le microphone considéré. Si les micros-bouches des deux pilotes ne sont pas sélectifs ou sont positionnés de telle sorte qu'ils captent toutes des sources émises dans le cockpit alors le mélange de SAS dans le CVR obtenu par fusion des Eqs. (2) et (3) résulte en un mélange convolutif. À l'inverse, si les micros-bouches des pilotes sont très sélectifs et ne captent que la voix des pilotes qui les portent, alors les valeurs de $A_{ij}(\omega)$ dans l'Eq. (3) sont négligeables pour $j \neq i$ et en notant $S'_i(\omega, n) \approx A_{ii}(\omega)S_i(\omega, n)$ l'image de la source i dans le microphone i, l'Eq. (2) se réduit à un mélange instantané des images $S'_i(\omega, n)$:

$$X_i(\omega, n) = S'_i(\omega, n) + \lambda_i S'_{3-i}(\omega, n). \quad (4)$$

Ce modèle théorique de mélange est valable pour les systèmes audio d'un très grand nombre de types d'avion mais va dans la pratique être confronté à de nombreuses sources de variabilité comme par exemple la géométrie des cockpits, la disposition et le volume sonore des haut-parleurs et des sources, le bon positionnement, la sélectivité et la sensibilité des micros-bouches, les positons relatives instantanées entre les sources et les microphones variant dans le temps en fonction notamment des mouvements de tête des pilotes. Ces configurations, qui peuvent également être dissymétriques entre les deux pilotes et varier au cours d'un même enregistrement CVR, définissent un système complexe dont le modèle de mélange se situe sur un point mobile entre un modèle convolutif et un modèle LI.

3.3 La SAS en support de l'activité de transcription CVR

L'évaluation de plusieurs méthodes de SAS est menée sur des enregistrements de CVR anonymisés d'événements non majeurs avec l'objectif d'évaluer leur bénéfice pour la tâche de transcription de segments de parole superposée partiellement inintelligibles. Nous faisons le choix ne n'utiliser que des méthodes de SAS dont les sources sont libres et dont les codes sources sont accessibles. Malgré le nombre croissant de méthodes fondées sur l'apprentissage profond nous ne les utiliserons pas dans cette évaluation car elles nécessitent des données d'apprentissage ce qui ne correspond pas à notre cadre d'utilisation. Nous avons retenu trois méthodes SAS qui s'appliquent toutes dans le domaine de représentation de la TFCT. La méthode DEMIX (Arberet *et al.*, 2010), utilisée ici dans sa version pour mélanges linéaires instantanés, cherche à estimer de manière robuste les coefficients de la matrice de mélange par une approche de clustering. La seconde méthode, UCBSS (Reju *et al.*, 2010), est fondée sur un algorithme de clustering développée pour les mélanges convolutifs. Les sources estimées

		DEMIX	UCBSS	NMF	3 méthodes
Pourcentage de segments améliorés	pilote / pilote	28%	16%	20%	44%
	pilote / radio	33%	28.5%	38%	66%
Taux de reconnaissance de mots	pilote / pilote	50%	55%	57.5%	80%
	pilote / radio	56%	63.6%	70%	89.6%

TABLE 1 – Performances obtenues après application de la SAS.

par DEMIX et UCBSS sont reconstruites par masquage binaire dans le domaine temps-fréquence. Nous évaluons aussi une méthode de NMF-SAS (Ozerov & Févotte, 2010) dans sa variante utilisant l'algorithme espérance-maximisation. Les sources y sont estimées par filtrage de Wiener.

Les enregistrements audio des CVR sont parmi des données les plus sensibles manipulées dans le contexte des enquêtes de sécurité. A l'instar de la *Circulaire n° 30053 du 3 septembre 2019 relative à l'export de données à destination de la recherche scientifique* établie par la Gendarmerie Nationale, le BEA encadre l'utilisation de ces données tout en encourageant des collaborations avec les chercheurs académiques. Le corpus utilisé dans cette évaluation se compose de 25 segments de parole pour le scénario considérant la parole superposée des pilotes et 21 segments de parole pour le scénario dans lequel la voix d'un pilote est couverte par la radio. Pêle-mêle, ces segments audio sont issus de 15 événements, contiennent les voix de 12 hommes et de 3 femmes, enregistrés par des CVR de 3 constructeurs distincts, installés sur 10 types d'avion différents et sur des phases d'opération variées (parking, roulage, décollage, croisière, approche, atterrissage).

Un analyste audio produit tout d'abord une transcription des segments du corpus d'évaluation en indiquant les termes incertains ou inintelligibles par un point d'interrogation (**?**). Les segments considérés comptent tous au moins un mot n'ayant pu être transcrit avec certitude par l'analyste. Les méthodes de SAS sont ensuite appliquées à ces segments. Lorsque la SAS apporte une amélioration en terme de séparation de sources, l'analyse audio se sert alors des signaux sources estimés pour améliorer la transcription initiale. Dans cette situation les performances de chaque méthode sont évaluées en termes de Taux de Reconnaissance de mots inintelligibles, qui est le pourcentage de mots initialement inintelligibles qui ont pu finalement être transcrits après application de la SAS.

Comme cela est présenté en Table 1, dans le scénario considérant les voix superposées des pilotes, l'intelligibilité des échanges est améliorée significativement par au moins une méthode de SAS sur seulement 44% segments audio traités. L'écoute conjointe des sources estimées par les trois méthodes permettent de transcrire avec confiance 80% des mots initialement inintelligibles. Comme indiqué en Table 1, nous observons que dans ce scénario, DEMIX est la méthode de SAS produisant les de sources estimées les plus intelligibles sur une majorité de segments. Dans le second scénario, l'intelligibilité est améliorée pour 66% segments considérés. L'utilisation de la SAS permet de transcrire 89.6% des termes initialement inintelligibles. La méthode de SAS fondée sur la NMF se montre plus performante vis-à-vis du nombre de mots reconnus et produit également dans ce scénario les sources estimées les plus intelligibles sur une majorité de segments audio.

4 Conclusion et perspectives

Ce travail est la toute première étude proposant d'utiliser des méthodes de Séparation Aveugle de Sources pour améliorer l'intelligibilité des séquences de parole superposée dans des enregistrements

de « boites noires » aéronautiques. L'évaluation de 3 méthodes de SAS sur un corpus d'enregistrements réels a révélé que la SAS est une solution capable dans certaines situations d'améliorer l'intelligibilité de 80% à 90% des mots initialement inintelligibles en fonction du scénario considéré. Toutefois ces résultats doivent être nuancés car, d'une part, pour atteindre ces résultats il est nécessaire de combiner les signaux de sorties de toutes les méthodes de SAS testées et, d'autre part il subsiste un assez grand nombre de segments audio sur lesquels la SAS n'a pas montré d'impact positif sur l'intelligibilité. Les perspectives de ces travaux s'orientent assez naturellement vers l'analyse des facteurs de variabilité des résultats observés et de l'évaluation de méthodes basées sur d'autres critères que ceux testés dont des approches basées sur de l'apprentissage automatique.

Références

ABRARD F. & DEVILLE Y. (2005). A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, **85**(7), 1389–1403.

ARBERET S., GRIBONVAL R. & BIMBOT F. (2010). A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Trans. on Signal Proc.*, **58**(1), 121–133.

P. COMON & C. JUTTEN, Eds. (2010). *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. Elsevier.

LEE D. & SEUNG S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**.

NUGRAHA A. A., LIUTKUS A. & VINCENT E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM TASLP*, **24**(9), 1652–1664.

OZEROV A. & FÉVOTTE C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE TASLP*, **18**(3), 550–563.

PUIGT M. (2007). *Méthodes de séparation aveugle de sources fondées sur des transformées temps-fréquence. Application à des signaux de parole.* Thèse de doctorat, Toulouse-III-Paul-Sabatier.

PUIGT M., VINCENT E. & DEVILLE Y. (2009). Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources. In *Proc. of ICA'09* : Springer-Verlag Berlin Heidelberg.

REJU V. G., KOH S. N. & SOON I. Y. (2010). Underdetermined convolutive blind source separation via time–frequency masking. *IEEE Trans. on Audio, Speech, and Language Processing*, **18**, 101–116.

SMARAGDIS P. & BROWN J. (2003). Non-negative matrix factorization for polyphonic music transcription. In *Proc. of WASPAA'03*, p. 177 – 180.

WATANABE S., MANDEL M., BARKER J., VINCENT E., ARORA A., CHANG X., KHUDANPUR S., MANOHAR V., POVEY D., RAJ D., SNYDER D., SHANMUGAM SUBRAMANIAN A., TRMAL J., BEN YAIR B., BOEDDEKER C., NI Z., FUJITA Y., HORIGUCHI S., KANDA N., YOSHIOKA T. & RYANT N. (2020). Chime-6 challenge : Tackling multispeaker speech recognition for unsegmented recordings. In *CHiME 2020 6th Intern. Workshop on Speech Processing in Everyday Environments*.

WILSON K. W., RAJ B., SMARAGDIS P. & DIVAKARAN A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *Proc. of ICASSP'08*, p. 4029–4032.

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