Apport de la séparation de sources au traitement des signaux audiophoniques issus de « boîtes noires aéronautiques »

Matthieu PUIGT[1], Benjamin BIGOT[2], Hélène DEVULDER[1,2]

[1]Univ. Littoral Côte d'Opale, LISIC – UR 4491, 62228 Calais, France

[2]Bureau d'Enquêtes et d'Analyses (BEA), 10 rue de Paris, 93352 Le Bourget, France

matthieu.puigt@univ-littoral.fr, benjamin.bigot@bea.aero

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 de la séparation de source afin d'améliorer le traitement des enregistrements CVR.

Abstract – BEA is the French authority in charge of safety investigations in case of a civil aircraft accident. In this context, BEA analyses audio from cockpit voice recorders (CVR). Constraints on CVR design result in the presence of a large amount of superimposed speech, which complicates CVR analysis. In this paper, we present a first study on the contribution of source separation for CVR recordings improvement.

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 – connus sous le nom de « boîtes noires » – comptant 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 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. La conception même des CVR amène à trouver, sur les 4 canaux audio enregistrés simultanément, une forte proportion de parole superposée. La parole superposée est également potentiellement plus fréquente lors d'incidents ou d'accidents durant lesquels l'activité vocale et sonore peuvent se densifier, ce qui complique le travail des analystes et peut 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) sur l'intelligibilité de la parole.

La SAS est un problème générique dont les premiers travaux ont été proposés en France au milieu des années 1980 [1]. 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. Dans cet article, nous nous intéressons à l'apport de la SAS, lorsque de multiples sources de parole sont superposées dans les enregistrements du CVR, pour les travaux de transcription, de segmentation et d'identification sonore réalisés au BEA.

Ce travail est à notre connaissance la première application des méthodes de SAS sur des enregistrements réels de CVR. L'article est structuré comme suit. Nous présentons en section 2 le système audio CVR et nous proposons un modèle des mélanges de sources sonores. Nous présentons en section 3 les résultats d'une étude préliminaire sur l'apport de la SAS pour les signaux CVR. Nous concluons avec les perspectives de ces travaux en section 4.

2 Modélisation du CVR

Un CVR 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. La réglementation définit le contenu des canaux enregistrés par les CVR embarqués dans les avions de transport commercial. Le canal n‘°1 (respectivement n°2) contient les signaux émis et reçus par le système audio du pilotes en place gauche (respectivement droite). Le canal n°3 contient les signaux émis et reçus par le système audio du poste $3^{\text{ème}}$ homme et les annonces aux passagers. Le canal n°4 correspond au microphone d'ambiance (CAM pour *Cockpit Area Microphone*). Le CAM est un microphone omnidirectionnel installé généralement sur le plafond du cockpit entre les pilotes. Ce canal, non considéré dans cette étude, capte les conversations et l'ambiance sonore du cockpit, dont les signatures spectrales des groupes motopropulseurs. Avant d'être enregistrés, les signaux des voies pi-

lotes et du CAM sont échantillonnés respectivement à 7 kHz et 12 kHz, et compressés au standard ADPCM.

Les trois canaux « pilote » du CVR contiennent chacun une combinaison des signaux reçus et émis à chaque poste, ce qui correspond concrètement à une superposition des sons entendus dans le casque et ceux captés par les microphones de chaque membre d'équipage. Les sources sonores disponibles dans les casques des pilotes sont typiquement les sons perçus par les microphones des autres postes d'équipage, les messages radio reçus du contrôle aérien et des autres aéronefs sur la fréquence, ainsi que les communications avec le personnel commercial. L'activation de ces sources ainsi que leurs niveaux sonores dans un casque sont ajustés par chaque pilote grâce à un panneau de réglage individuel. Chaque poste pilote est équipé d'un casque-micro, d'un microphone à main et d'un troisième microphone monté à l'intérieur d'un masque à oxygène. Ces microphones captent surtout la voix du pilote qui les utilise, 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. Les pilotes n'entendent pas leur propre voix dans leur casque, sauf lorsqu'ils émettent sur le canal radio. Les signaux envoyés dans le casque d'un pilote sont également reproduits par des haut-parleurs situés à l'avant gauche et droit du cockpit.

Les signaux enregistrés par le CVR sont légèrement différents de ceux disponibles aux postes de pilotage. D'une part les niveaux respectifs de chaque source de l'écoute présentée au CVR sont réglés et figés lors de l'installation du CVR et ne reflètent pas les ajustements individuels des pilotes D'autre part, à la demande des enquêteurs de sécurité, une fonction « hot mic » ou « microphone ouvert » est implémentée sur les micros-bouche et les microphones des masques à oxygène. Enfin, lors de leur superposition avant enregistrement par le CVR, les niveaux relatifs des signaux provenant des microphones et de ceux provenant de l'écoute d'un pilote sont adaptés dynamiquement par le système audio de l'avion de manière à garantir une certaine intelligibilité de la parole.

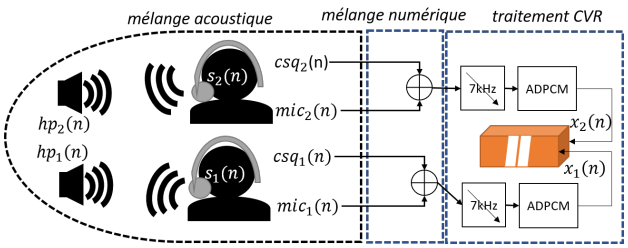

FIGURE 1 – Mélange de sources sonores dans le CVR.

Dans cet article, nous proposons le premier modèle de mélange dédié aux CVR. L'originalité de ce mélange réside dans son côté hybride, avec un mélange acoustique suivi d'un mélange numérique et d'une étape de compression (voir figure 1). Pour des raisons de place, nous définissons ce modèle dans le domaine temps-fréquence obtenu après application d'une transformée de Fourier à court terme (TFCT) des signaux. Nous considérons les sources d'intérêt notées $S_1(\omega, n)$ et $S_2(\omega, n)$ qui correspondent aux signaux de parole des pilotes en place gauche et en place droite. Ces signaux se propagent de manière acoustique jusqu'aux microphones $MIC_i(\omega, n)$ ($i \in \{1, 2\}$) et sont notamment mélangés avec les signaux $HP_i(\omega, n)$ ($i \in \{1, 2\}$) fournis par les haut-parleurs situés généralement vers l'avant à gauche et droit du cockpit. L'ensemble de cette chaîne de propagation des sources sonores peut être modélisée selon un modèle convolutif, exprimé ici dans le domaine fréquentiel :

$$\begin{aligned} MIC_i(\omega, n) \approx\ & A_{i1}(\omega)S_1(\omega, n) + A_{i2}(\omega)S_2(\omega, n) \\ & + A_{i3}(\omega)HP_1(\omega, n) + A_{i4}(\omega)HP_2(\omega, n), \end{aligned} \quad (1)$$

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 (pilote ou haut-parleur) vers le microphone considéré.

Les signaux joués par les haut-parleurs consistent en un mélange que nous supposons linéaire instantané (LI). En particulier, le signal $HP_i(\omega, n)$ ($i \in \{1, 2\}$) s'écrit :

$$HP_i(\omega, n) = \alpha_{i1} CSQ_i(\omega, n) + \alpha_{i2} ALM(\omega, n), \quad (2)$$

où $CSQ_i(\omega, n)$ est la signal entendu dans le casque du pilote i, $ALM(\omega, n)$ est l'ensemble des alarmes qui sonnent dans le cockpit, α_{i1} et α_{i2} sont les coefficients de mélanges. Alors que α_{i1} est réglé manuellement par le pilote i (pour lui permettre d'entendre ce qui passe dans son casque même s'il ne le porte pas), la valeur de α_{i2} est fixée automatiquement par l'avion selon la phase de vol et est la même dans les deux haut-parleurs.

Les signaux joués dans les casques sont eux-aussi modélisés comme des mélanges LI de plusieurs signaux. En particulier, le pilote i entend dans son casque les signaux suivants :

$$CSQ_i(\omega, n) = \beta_{i1} MIC_{3-i}(\omega, n) + \beta_{i2} R(\omega, n), \quad (3)$$

où $R(\omega, n)$ est le canal radio et β_{ij} ($j \in \{1, 2\}$) sont les coefficients de mélanges réglés manuellement par le pilote.

En combinant les équations (1), (2) et (3), on obtient des boucles : par exemple $MIC_1(\omega, n)$ capte le signal $S_1(\omega, n)$ directement mais aussi à travers $HP_2(\omega, n)$. En pratique, les niveaux sonores sont réglés de manière à éviter tout larsen.

Enfin, les signaux enregistrés dans les CVR correspondent à des mélanges des signaux provenant du microphone et des sons activés dans le casque de chaque pilote, à nouveau selon un mélange que nous supposons LI et préréglé lors de l'installation de l'enregistreur :

$$X_i(\omega, n) = \gamma_{i1} MIC_i(\omega, n) + \gamma_{i2} MIC_{3-i}(\omega, n) + \gamma_{i3} R(\omega, n). \quad (4)$$

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 avec en particulier, 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-bouche, les positions relatives instantanées entre les sources et les microphones variant dans le temps en fonction notamment des mouvements de tête des pilotes. De plus, ces caractéristiques peuvent ne pas être symétriques entre le pilote en place

gauche et celui en place droite. Toutes ces sources de variabilité vont avoir pour conséquences de positionner le modèle de mélanges réel d'un enregistrement CVR, voire d'un segment de parole, quelque part entre un modèle convolutif dynamique et un modèle LI plus ou moins complexe. De plus l'exemple fourni ici ne concerne que des mélanges où seuls deux membres de l'équipage sont présents.

En effet, en combinant les équations (1) à (4), on obtient un mélange globalement convolutif des sources de parole de l'équipage, des signaux de la radio et des alarmes mais ce mélange peut être simplifié dans de nombreuses situations. En particulier, si $ALM(\omega, n) = 0$ et si le microphone i est très sélectif (notamment, s'il est bien positionné en face de la bouche du pilote), l'équation (1) se simplifie puisqu'on peut alors supposer que $A_{i,3-i}(\omega)$, $A_{i,4-i}(\omega)$ et $A_{i,5-i}(\omega)$ sont négligeables pour toute pulsation ω. Dans ce cas très précis, il en résulte que

$$MIC_i(\omega, n) \approx A_{ii}(\omega) S_i(\omega, n) \triangleq S'_i(\omega, n), \qquad (5)$$

et les mélanges audios des CVR peuvent être perçus comme des mélanges LI :

$$X_i(\omega, n) \approx \gamma_{i1} S'_i(\omega, n) + \gamma_{i2} S'_{3-i}(\omega, n) + \gamma_{i3} R(\omega, n). \quad (6)$$

Au contraire, si les micros-bouches sont peu sélectifs, les signaux émis par les hauts-parleurs sont alors généralement captés par les deux microphones et on obtient un mélange hybride avec des combinaisons LI des signaux $S'_i(\omega, n)$ et des mélanges convolutifs des autres sources sonores $R(\omega, n)$ et $ALM(\omega, n)$.

Nous tenons à préciser que les modèles présentés dans cet article ont été obtenus par rétro-ingénierie des enregistrements CVR et sur la base d'échanges avec des pilotes. Les connaissances précises de chaque système relèvent du secret industriel.

3 Application de la SAS aux CVR

Nous choisissons des méthodes de SAS dont les sources sont libres et dont les codes sources sont accessibles. De plus, compte tenu de la multiplicité des scénarios de mélanges possibles allant du cas sur-déterminé au sous-déterminé, nous retenons trois méthodes d'analyse en composantes parcimonieuses [1, Ch. 10] dans cette étude préliminaire. En particulier, nous avons choisi DEMIX [2], DUET [3] et UCBSS [4] respectivement développées pour les mélanges LI, anéchoïques et convolutifs. Une fois les paramètres de mélanges estimés, les sources sont obtenues par masquage [1] binaire dans le domaine temps-fréquence [3].

Le corpus d'évaluation compte 18 segments de parole issus de 12 enregistrements CVR. Ces extraits audio contiennent les activités superposées de 2 à 4 sources sonores correspondant aux activités vocales des pilotes en poste, à des communications entrantes sur la radio, à des annonces aux passagers et à des alarmes sous forme d'annonce par voix synthétique. Le nombre de segments évalués dans notre étude est relativement

1. D'autres méthodes de reconstruction théoriquement plus performantes ont aussi été testées dans des analyses préliminaires mais n'ont pas fourni de résultat exploitable. Elles ne sont donc pas utilisées dans cette étude.

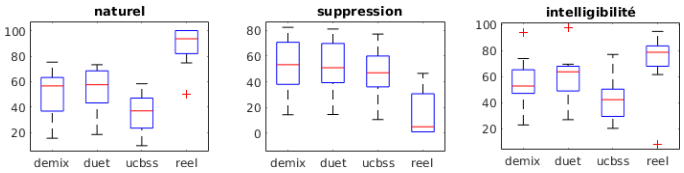

FIGURE 2 – Résultats de l'évaluation subjective de la qualité.

important, ce qui est cohérent avec l'importante diversité de situations observées dans des enregistrements CVR réels.

L'évaluation subjective des performances de SAS, telle que décrite dans [5], est mise en oeuvre avec une légère adaptation due au fait que nous n'ayons pas accès aux sources de références. Finalement nous proposons à une cohorte de 11 auditeurs, parmi lesquels 3 spécialistes des données CVR, d'évaluer la qualité des sources estimées selon les 3 critères suivants :

— l'intelligibilité globale de la source reconstruite,
— l'absence de bruit musical dans la source reconstruite, et
— la suppression des autres sources dans les sorties.

Pour chacun des 18 segments du corpus, pour éviter une durée excessive des tests dans un contexte sanitaire particulier lié au retour de confinement, une seule source estimée à l'aide de chacune des méthodes ainsi qu'une seule observation non-traitée du CVR sont présentées aux auditeurs qui évaluent à chaque passe un seul des critères de performance. L'évaluation est réalisée grâce à l'interface graphique disponible dans la boîte à outils PEASS [5]. Au début de chaque phase de test, des instructions propres au critère évalué sont fournies aux auditeurs qui devront par la suite attribuer un score entre 0 et 100 pour chaque signal entendu. Après un court entraînement permettant d'écouter des exemples sonores illustrant le critère de qualité visé, les 18 jeux de 4 signaux à évaluer sont présentés successivement dans un ordre aléatoire. L'évaluation dure environ 1 heure et 15 minutes. Une pause de 15 minutes est imposée entre chaque passe.

La figure 2 présente les résultats obtenus par les 11 auditeurs sur l'ensemble du corpus pour chacune des méthodes de SAS et pour chacun des critères de qualité considérés. Les boites à moustaches représentent l'intervalle de confiance à 95%, la médiane et les valeurs maximales et minimales des scores attribués. Les valeurs aberrantes y sont représentées par des croix.

Les scores attribués aux sources estimées par les méthodes DUET et DEMIX sont très similaires sur les trois critères évalués. L'algorithme UCBSS obtient des scores moins importants mais souffre d'un fort bruit musical introduit lors de la reconstruction, ce qui explique de moins bons scores d'intelligibilité et de sonorité naturelle. Le test a été jugé difficile par une majorité des auditeurs non-experts et, dans ce sens, les intervalles de confiance à 95% montrent un taux d'accord inter-annotateur plus important sur les signaux non traités que pour les signaux issus de la SAS pour les critères d'intelligibilité et sur l'absence de bruits musicaux (critère titré « naturel » en figure 2). La reconstruction par masquage temps-fréquence ajoute en effet une distorsion et un bruit musical aux signaux reconstruits. La capacité des méthodes de SAS à séparer la source prin-

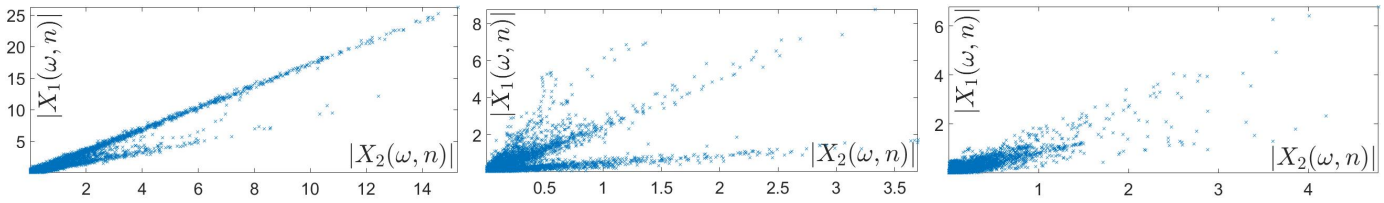

FIGURE 3 – Nuages de points, avec de g. à d. : mélange LI selon l'Eq. (6), mélange hybride LI/convolutif, mélange convolutif.

cipale des autres sources, titré « suppression » en figure 2, a globalement été considéré efficace et satisfaisante par une majorité d'annotateurs, quelle que soit la méthode de SAS appliquée. Cependant, cette évaluation ne montre pas d'amélioration significative de l'intelligibilité de la source principale. Les auditeurs non-experts ont même considéré que l'intelligibilité était dégradée par l'application des algorithmes de SAS. Contrairement à ce résultats, les trois spécialistes des données CVR ont tous considéré que l'intelligibilité avait été améliorée. Après discussion avec les évaluateurs, il ressort que les scores d'intelligibilité produits par les non-experts doivent être considérés avec des réserves car une majorité d'auditeurs ont évalué l'intelligibilité en incluant également les résidus de sources présents dans les signaux reconstruits au lieu de se concentrer uniquement sur l'intelligibilité de la source principale. Les spécialistes des données CVR, plus habitués à discriminer les sources sonores « à l'oreille » se sont naturellement concentrés sur la source principale. L'accord inter-évaluateur est également beaucoup plus important sur les scores attribués par les 3 spécialistes. Les travaux en cours se focalisent uniquement sur l'intelligibilité, dans des protocoles plus simples [6].

Pour illustrer la grande variabilité des mélanges disponibles dans le corpus d'évaluation, nous présentons figure 3 le nuage de points du module de la TFCT $X_1(\omega, n)$ en fonction du module de $X_2(\omega, n)$. Le nuage de points représenté sur la figure de gauche montre des points concentrés sur deux directions bien définies (mélange parole/radio). Ce type de nuage de points est caractéristique de signaux disjoints aisément séparables par masquage temps-fréquence. La figure du milieu laisse apparaître trois directions avec une dispersion plus importante du nuage de points autour des directions, ce qui semble indiquer un mélange plus réverbérant de sources parcimonieuses (parole / annonce aux passagers). Le nuage de point représenté sur la figure de droite ne permet pas d'isoler visuellement une direction particulière, ce qui correspond à un mélange plus complexe de signaux parcimonieux (parole / alarmes).

4 Conclusion et perspectives

Nous proposons dans ce papier une première étude sur l'apport de la séparation aveugle de sources sur des enregistrements réels d'enregistreurs de conversations, plus communément appelés « boîtes noires ». Les contraintes de conception de ces dispositifs conduisent des enregistrements à contenir une proportion non négligeable de parole superposée avec un impact potentiellement négatif sur l'exploitation des contenus audio.

Dans cette étude, nous présentons un modèle théorique de mélange des sources sonores présentes dans ces enregistrements et mettons en avant la grande variabilité existante autour de ce modèle. Nous réalisons également une évaluation subjective sur un corpus composé de 18 extraits issus d'enregistreurs réels. Les évaluations menés par une cohorte de 11 auditeurs spécialistes et non-spécialistes de ces données atypiques ont montré des résultats encourageants sur la capacité des algorithmes évalués à améliorer ces enregistrements en termes d'intelligibilité et de suppression de sources.

Les perspectives de ces travaux nous orientent vers l'analyse des facteurs de variabilité des mélanges et l'évaluation des performances sur la base d'autres critères que ceux testés ici. Dans ce but nous avons récemment présenté [6] une seconde étude axée sur l'amélioration de la qualité de transcription de la parole et avons considéré d'appliquer la séparation de source à des mélanges de signaux plus limités mais mieux contrôlés. Cette approche pourrait être un point d'entrée pour une utilisation efficace de la séparation de sources sur ces enregistrements difficiles dans un cadre opérationnel.

Références

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation : Independent Component Analysis and Applications*, Elsevier, 2010.

[2] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, pp. 101–116, 2010.

[5] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[6] B. Bigot, H. Devulder, and M. Puigt, "Amélioration de l'intelligibilité de la parole dans des enregistrements de « boîtes noires aéronautiques » à l'aide de méthodes de séparation aveugle de sources," in *Actes des JEP*, juin 2022.