Introducing the "Cockpit Party Problem": Blind Source Separation Enhances Aircraft Cockpit Speech Transcription

Matthieu Puigt, Benjamin Bigot, Hélène Devulder

Paper reference: M. Puigt, B. Bigot, and H. Devulder, "Introducing the "Cockpit Party Problem": Blind Source Separation Enhances Aircraft Cockpit Speech Transcription," in Journal of the Audio Engineering Society, vol. 73, no. 1/2, pp. 43–53, 2025.

AES online version: `https://dx.doi.org/10.17743/jaes.2022.0189`.

Introducing the "Cockpit Party Problem": Blind Source Separation Enhances Aircraft Cockpit Speech Transcription*

MATTHIEU PUIGT,[1] (matthieu.puigt@univ-littoral.fr)
BENJAMIN BIGOT,[2] (benjamin.bigot@protonmail.com)
AND HÉLÈNE DEVULDER[1,2] (helene.devulder@free.fr)

[1] *Univ. Littoral Côte d'Opale, LISIC – UR 4491, F-62228 Calais, France*
[2] *BEA – Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, F-93350 Le Bourget, France*

Cockpit Voice Recorders (CVRs) are one of the two mandatory flight recording devices embarked in commercial aircraft. Its analysis is crucial to understand the context of an air incident or accident. However, in such scenarios, when the audio recordings are usable, CVR may contain strong mixtures of crew member speech signals, radio communications, and cockpit alarms. However, contrary to the "cocktail party problem" that Blind Source Separation (BSS) aims to tackle, modeling CVR mixtures—that we here name the "cockpit party problem"—was never done before. In this paper, we thus propose a CVR mixture model and highlights its limitations. While not trivial—even in a two-source scenario—BSS methods can be applied to real CVR recordings. We find that taking into account several BSS outputs provided by various methods may help audio analysts to transcribe the CVR data. That is near 90% of unintelligible words can be transcribed from CVR recordings processed by BSS methods.

0 INTRODUCTION

Public transportation aircraft are fitted with two crash-survival flight recorders—also known as "black boxes"—which are named the Cockpit Voice Recorder (CVR) and the Flight Data Recorder, and which need to be retrieved and analysed by air accident authorities in case of incident or accident. BEA (*Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile*) is the French authority in charge such investigations. CVR contents are "manually" transcribed by specialised investigators (a.k.a. audio analysts) for the benefits of the safety investigation (Fig. 1).

In a CVR recording, the causes of speech intelligibility degradation are numerous. In particular, the CVR design itself generates a significant amount of superimposed—a.k.a. mixed—speech signals over the 4 audio channels which are simultaneously recorded. Moreover, in case of an aircraft accident or incident, superimposed speech signals are more likely to occur—since voice and cockpit sound activities become denser—which may yield to the loss of crucial information for the safety investigators. BEA already uses sound source subtraction algorithms and aims to investigate the enhancement provided by Blind Source Separation (BSS) on CVR speech intelligibility.

BSS is a generic problem whose pioneering work emerged 40 years ago [1]. When applied to audio signals, such a problem is also known as the "cocktail party problem" [2] and aims to unmix N unknown audio sources from M observation signals which are obtained from distant microphones, such that each microphone recording contains mixtures of these source signals. In this paper, we investigate the BSS enhancement in CVR recordings to help the BEA audio analysts in their sound transcription, segmentation, and identification tasks. To the best of the authors' knowledge, such a work is the very first one to be conducted on real CVR recordings. The overall contribution of the paper is two-fold. Firstly, we propose a CVR mixing model for the pilots' channels. We then discuss its validity and its limitations. Secondly, we consider some situations when classical BSS methods can be applied and we investigate how BSS outputs can help the BEA audio analysts in their tasks.

The remainder of the paper is organised as follows. We briefly recall the history of BSS in Sect. 1. We introduce the CVR audio system in Sect. 2, for which we propose a dedicated sound source mixture model. We investigate in Sect. 3 the enhancement provided by some state-of-the-art BSS methods on CVR transcription. We conclude and discuss about future work in Sect. 4.

*To whom correspondence should be addressed, e-mail: matthieu.puigt@univ-littoral.fr. Last updated: August 5, 2024

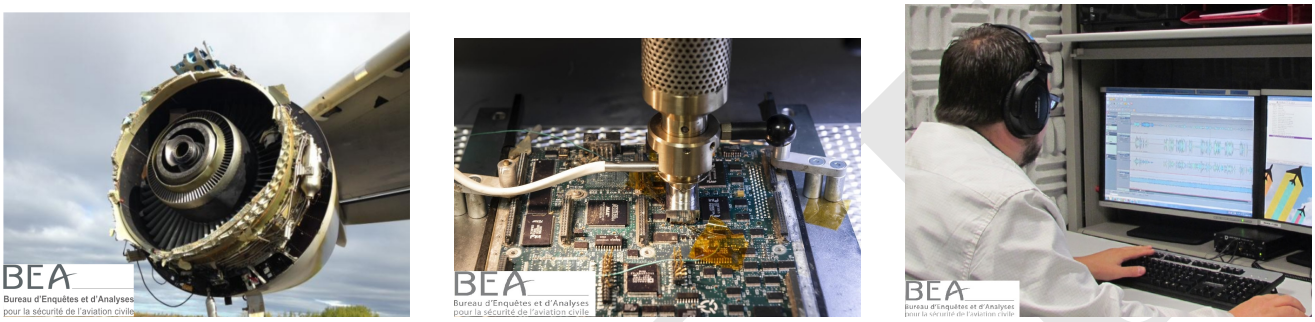

Fig. 1. After a civil aircraft incident / accident (left), the CVR is opened (middle), and an audio analyst transcribes its content (right).

1 RELATED WORK

CVR recordings are not really known in the scientific community. In order to properly define them in Section 2, we first briefly recall the state-of-the-art in BSS.

1.1 Mixture Models

BSS is a generic problem which has been investigated for many applications [1]. When applied to audio signals, BSS aims to solve the well-known "cocktail party problem" [2]. In that framework, we assume that N unknown acoustic sources are active and are recorded by M microphones which provide mixtures of these sources. Modeling the acoustic propagation have been intensively investigated in the pioneering works on BSS [1]. In particular, many investigations focused on linear mixtures, i.e., the linear instantaneous (LI), the anechoic and the convolutive mixtures (see Fig. 2).

The LI one can model mixtures from a multitrack recording (provided no additional filtering effect is added). Denoting $x_i(n)$, $s_j(n)$, and a_{ij} the i-th observation, the j-th source signal, and the contribution of Source j in Observation i, respectively, the observation signals following the LI model read

$$\forall i = 1,\dots,M, \quad x_i(n) = \sum_{j=1}^{N} a_{ij} s_j(n). \tag{1}$$

Due to its simplicity, the LI model has been extensively investigated, so that the community concluded that "the mixing matrix estimation task is now solved for instantaneous mixture" [3].

The anechoic mixture model takes into account the direct propagation time of the acoustic wave from each source to each microphone. Denoting δ_{ij} the time shift related to Source j in Observation i, this model reads

$$\forall i = 1,\dots,M, \quad x_i(n) = \sum_{j=1}^{N} a_{ij} s_j(n-\delta_{ij}). \tag{2}$$

This model has been comprehensively investigated in BSS—e.g., in [4, 5, 6, 7]—and was also considered for BSS-inspired audio source localisation methods [8, 9].

Lastly, the convolutive mixture is the more general linear model. It assumes that each microphone receives several attenuated and time-shifted versions of each source, thus resulting in a filtered version of the latter. Denoting $a_{ij}(n)$ the propagation filter from Source j to Microphone i and $\star$ the convolution operation, the observations read

$$\forall i = 1,\dots,M, \quad x_i(n) = \sum_{j=1}^{N} a_{ij}(n) \star s_j(n). \tag{3}$$

Such linear models have been extended in several ways. First of all, they are here defined for fixed source and microphone positions. However, when the sources and/or the microphone move, the mixing parameters evolve with time as well. Then, some authors started to handle nonlinear effects such as microphone saturation through the post-nonlinear (convolutive) mixture model [10, 11, 12] or the clipped mixture model [13]. Lastly, the lossy audio coding effect is usually not taken into account while it may have a major effect on the reached BSS performance [14][1].

1.2 History of the BSS Methods

In terms of methods, the historical approaches are based on Independent Component Analysis (ICA) [1]. They aim to combine the observations so that the ICA outputs are statistically independent. Assuming that the true sources $s_j(n)$ are independent as well, the ICA outputs are then equal to the sources, up to permutation and scale/filter ambiguities [16]. These approaches were then generalised under the Independent Vector Analysis (IVA) framework [17]. When applied to convolutive BSS problem, IVA allows to solve the permutation ambiguity which occurs in frequency-domain ICA [18]. As an alternative to independence-based approaches, some methods based on other source assumptions were proposed since the end of the 90s.

On one side, approaches based on Time-Frequency (TF) representations and source sparsity emerged. While some authors investigated the use of quadratic TF transformations and proposed methods inspired by second-order statistics BSS [19, 20], most authors focused on atomic TF decompositions and investigated techniques named Sparse Component Analysis (SCA) [21]. These approaches allow to separate sources in both (over-)determined mixtures—i.e., when the number N of sources is (strictly) less than the number M of microphones—and under-determined mixtures, i.e., when $N > M$. Some SCA methods, e.g., [22, 23], also allow to separate statistically dependent sources.

[1] However, some *informed* source separation methods were specifically proposed to tackle that issue, e.g., [15].

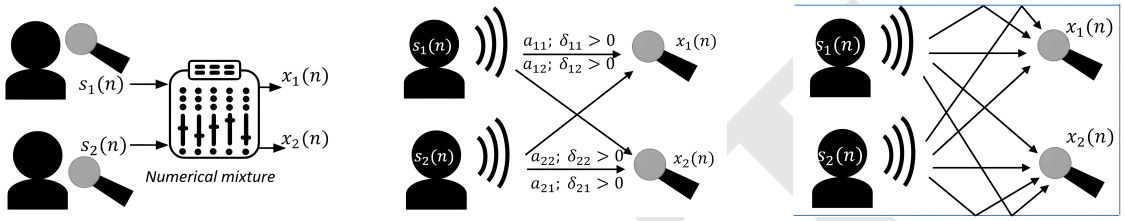

Fig. 2. Classical linear mixing models met in audio source separation.

The first SCA methods assumed that in any TF point [5] or in some TF areas to find [22, 23, 7], only one source is active. These assumptions were further relaxed, allowing strictly less than M sources to be active in each TF point [24, 25] or that some sources are unactive in some areas to find [26].

On the other side, methods based on Nonnegative Matrix Factorization (NMF) and Nonnegative Tensor Factorization (NTF) [27] became extremely popular. NMF is usually applied to decompose a spectrogram as the product of two nonnegative matrices, one of them corresponding to spectral signatures while the second one is a matrix of time-activation weights [28]. NMF quickly became the state-of-the-art in audio BSS, especially as it could be performed in both a supervised way—i.e., one matrix being trained with clean signals—and an unsupervised way. Some of the major extensions of NMF include multichannel NMF [29] or Deep NMF [30]. However, these approaches need the phase information to be estimated (as it is lost prior to the decomposition) and alternative were proposed, e.g., complex NMF [31]. NTF can be seen as a "natural" multichannel extension of NMF and several authors investigated its use for BSS, e.g., [32, 33]. Lastly, please note that there exist some similarities between NMF and SCA [34].

More recently, techniques based on deep learning emerged. While some approaches could be seen as deep extensions of supervised NMF/NTF techniques [35], there is a tremendous effort to develop novel methods with different network architectures, e.g., combinations of convolutional neural networks and long-short term memory [36], or attention-based networks [37]. We invite interested readers to discover several surveys on this topic [38, 39].

1.3 Discussion

The above mixing models and the state-of-the-art approaches were proposed for classical audio settings. However, to the best of the authors' knowledge, no existing work consider BSS on real CVR signals. Indeed, the authors in [40] claim to investigate the use of ICA for CVR signals. However, they consider a "simulation" where one source is an alarm while the other one is some background noise. These sources are recorded in a true cockpit using a mobile recording system, and are not issued from a CVR. However, there exists some prior work on BSS for aircraft mechanical noise separation / enhancement [41, 42].

2 CVR MODELING

2.1 General Principles of the CVR

CVR is a 4-channel audio recording device. While they were previously recorded on magnetic bands until the end of 90s, CVR data are now digitised and saved on a memory card within a crash-survival box. The regulations define the content of the channels recorded by the CVRs on board commercial transport aircraft.

Channels 1 and 2 contain the signals which were emitted and received by the audio system of the pilots in the left and right seats, respectively. Channel 3 contains signals transmitted and emitted by the audio system of the third crew position (jump-seat) and the messages to the passengers. Channel 4 corresponds to the Cockpit Area Microphone (CAM). CAM is an omnidirectional microphone usually installed on the cockpit ceiling between the pilots (see Fig. 3). The CAM channel records the several sounds and the speech communications in the cockpit. It also captures some spectral content about the aircraft powertrains. However, as it is sampled at 12 kHz—while the other CVR signals are sampled at 7 kHz—we do not consider it in this study. Lastly, all the CVR signals are filtered and coded using a lossy audio compression as the Adaptive Differential Pulse Code Modulation (ADPCM).

Each of the three CVR "crew" channels contains a combination of signals which are received and emitted in each pilot headset, i.e., a mixture of sounds heard in the headset and recorded by the microphone of each crew member. While the aircraft can host a third pilot, such a situation only appears in a small number of flights and almost never happens in BEA investigations. Moreover, if such a scenario might be met, in practice, the third CVR channel is configured to record announcements to passengers by flight attendants and pilots, and significantly differs from the other CVR channels. As a consequence, we do not consider the third channel in the remainder of the paper.

Typically, the sound sources heard in a pilot headset are those recorded by the microphones of the other crew members, the radio messages received from air traffic control and other aircraft on the frequency, as well as communications with flight attendants. The activation of these sources as well as their sound levels in a headset are adjusted by each pilot thanks to an individual control panel. Each pilot seat is equipped with a headset, a handheld microphone, and a third microphone mounted inside an oxygen mask. Each of these microphones mainly picks up the voice of the pilot who uses it, but it is common to also perceive at a lower level the audio environment of the cockpit and


Fig. 3. Examples of a CVR system (left), a CAM (center), and a pilot headset (right).

in particular the sound alerts emitted by the loudspeakers of the cockpit. The pilots do not hear their own voice in their headset, except when transmitting on the radio channel. The signals sent into a pilot's headset are also reproduced through speakers located forward left and right of the cockpit.

However, the signals which are stored in the CVR system are slightly different from those available for the crew members. Indeed, the respective levels of each listening source presented to the CVR are adjusted during the installation of the CVR and do not faithfully reflect the individual adjustments of the pilots. Moreover, due to the request of safety investigators, a "hot mic" or "open microphone" function is implemented on mouth mics and oxygen mask microphones. Lastly, when they are superimposed before being recorded by the CVR, the relative levels of the sound signals recorded by the microphones and those heard in the pilot headsets are dynamically adapted by the aircraft audio system in order to guarantee a certain speech intelligibility.

2.2 CVR Mixture Model

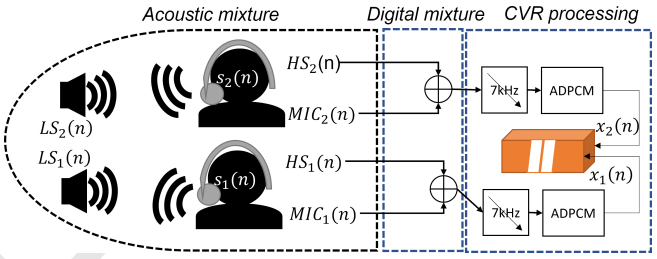

Fig. 4. Cockpit sound mixing process in the CVR.

In this paper and to the best of the authors' knowledge, the first mixing model dedicated to CVR recordings is proposed. We would like to point out that the models presented in this article were obtained by reverse engineering of CVR recordings and on the basis of exchanges with pilots. The precise knowledge of each system is an industrial secret. The novelty of such a model lies in its hybrid form as it consists of an acoustic mixture, followed by a numerical one and a compression stage (see Fig. 4).

For space considerations, we express it in the time-frequency domain obtained after a Short-Time Fourier Transform (STFT) of the signals. We consider the sources of interest—denoted $S_1(\omega,n)$ and $S_2(\omega,n)$, where ω and n stand for the angular frequency and time index of the considered time-frequency bin, respectively—which correspond to the speech signal of the pilots on the left and right seats, respectively. These signals are acoustically propagated to the microphones—whose recorded signals are denoted $MIC_i(\omega,n)$ $(i=1,2)$—and are mixed with acoustic signals emitted by the left and right cockpit loudspeakers denoted $LS_1(\omega,n)$ and $LS_2(\omega,n)$, respectively. The whole source propagation channel can be modeled by convolutive mixtures. Using the narrow-band frequency BSS approximation, they read

$$\begin{cases} MIC_1(\omega,n) \approx A_{11}(\omega)S_1(\omega,n) + A_{12}(\omega)S_2(\omega,n) \\ \qquad\qquad +A_{13}(\omega)LS_1(\omega,n) + A_{14}(\omega)LS_2(\omega,n), \\ MIC_2(\omega,n) \approx A_{21}(\omega)S_1(\omega,n) + A_{22}(\omega)S_2(\omega,n) \\ \qquad\qquad +A_{23}(\omega)LS_1(\omega,n) + A_{24}(\omega)LS_2(\omega,n), \end{cases} \tag{4}$$

where $A_{ij}(\omega)$ $(i=1,2,\ j=1,\ldots,4)$ is the Fourier transform of the acoustic propagation filter from a sound source (emitted by a pilot or a loudspeaker) to the i-th microphone.

Moreover, the signals played by both loudspeakers are themselves some mixtures of different sound sources. In particular, we assume that such a mixture is linear instantaneous (LI) and that both signals read

$$\begin{cases} LS_1(\omega,n) = \alpha_{11}HS_1(\omega,n) + \alpha_{12}ALM(\omega,n), \\ LS_2(\omega,n) = \alpha_{21}HS_2(\omega,n) + \alpha_{22}ALM(\omega,n), \end{cases} \tag{5}$$

where $HS_i(\omega,n)$ is the signal heard in the headset of Pilot i, $ALM(\omega,n)$ is the set of alarms which are activated in the cockpit, and α_{i1} and α_{i2} are the LI mixing coefficients. While α_{i1} is manually set by Pilot i—such that sounds played into its headset can be heard even if the latter is not worn—the value of α_{i2} is automatically fixed by the plane system with respect to the flight phase, and is set so that $\alpha_{12}=\alpha_{22}$. Furthermore, the signals played in the headsets are also modeled as LI mixtures of several signals, i.e.,

$$\begin{cases} HS_1(\omega,n) = \beta_{11}MIC_2(\omega,n) + \beta_{12}R(\omega,n), \\ HS_2(\omega,n) = \beta_{21}MIC_1(\omega,n) + \beta_{22}R(\omega,n), \end{cases} \tag{6}$$

where $R(\omega,n)$ is the radio channel and the β_{ij} coefficients denote the mixing parameters which are set by Pilot i.

Combining Eqs. (4), (5), and (6) yields some loops. For example, $MIC_1(\omega,n)$ directly captures $S_1(\omega,n)$ but also through $LS_2(\omega,n)$. In practice, the sound levels are set so that there is no feedback effect.

Lastly, the CVR channel associated with one pilot—say Pilot i—corresponds to a mixture of the signals recorded in the pilot microphone and of those heard in his headset. Again, we model such a mixture as an LI one, whose mixing coefficients are preset during the CVR installation in the cockpit aircraft, i.e.,

$$\begin{cases} X_1(\omega,n) = \gamma_{11} MIC_1(\omega,n) + \gamma_{12} MIC_2(\omega,n) + \gamma_{13} R(\omega,n), \\ X_2(\omega,n) = \gamma_{21} MIC_1(\omega,n) + \gamma_{22} MIC_2(\omega,n) + \gamma_{23} R(\omega,n). \end{cases} \quad (7)$$

Let us stress again that this model was obtained by reverse engineering of CVR recordings and by exchanging with pilots. Still, we analyse it and propose scenarios where it can be simplified below.

2.3 Analysis and Limitations

The proposed mixing model in Eq. (7) is valid for audio systems of a large amount of aircraft types. However, in practice, it will face many sources of variability, e.g., (i) the geometry of the cockpits, (ii) the layout and the sound volume of the loudspeakers and the sources, (iii) the correct positioning, the selectivity, and the sensitivity of the mouth microphones, (iv) the instantaneous relative positions between the sources and the microphones which vary over time with respect to pilot head movements. Moreover, these characteristics may not be symmetrical between the pilots in the left and right seats, respectively. All these sources of variability will make the real mixture model of a CVR recording—or even of a speech segment—somewhere between a dynamic convolutive model and a fixed or dynamic LI model.

Indeed, incorporating Eqs. (4) and (5) into Eq. (7), we obtain an overall convolutive mixture of speech sources, radio signals, and alarms but such a model may be simplified in many situations. More specifically, if $ALM(\omega,n) = 0$ in Eq. (5) and if Microphone i is very selective—in particular, if it is well placed in front of the pilot mouth—one may assume that in Eq. (4), $\forall i \in \{1,2\}$, $A_{i,3-i}(\omega)$, $A_{i,3}(\omega)$, and $A_{i,4}(\omega)$ are negligible over all the angular frequencies ω. In that case, Eq. (4) reads

$$\forall i = 1,2, MIC_i(\omega,n) \approx A_{ii}(\omega) S_i(\omega,n) \triangleq S'_i(\omega,n), \quad (8)$$

and the audio mixtures in the CVR recordings can be seen as LI mixtures of the pseudo-sources $S'_i(\omega,n)$ and of the radio signal $R(\omega,n)$, i.e.,

$$\begin{cases} X_1(\omega,n) \approx \gamma_{11} S'_1(\omega,n) + \gamma_{12} S'_2(\omega,n) + \gamma_{13} R(\omega,n), \\ X_2(\omega,n) \approx \gamma_{21} S'_1(\omega,n) + \gamma_{22} S'_2(\omega,n) + \gamma_{23} R(\omega,n). \end{cases} \quad (9)$$

On the contrary, if the microphones are not very selective, the signals emitted by the loudspeakers may then be picked up by both microphones and a hybrid mixture is obtained with LI combinations of the signals $S'_i(\omega,n)$ and convolutive mixtures of the other sound sources, i.e., $R(\omega,n)$ and $ALM(\omega,n)$.

It is worth noting that the above signals are sparse in the time-frequency domain [5], so that the scatter plots of their LI mixtures—i.e., the plot obtained by drawing the modulus of the TF transform of one observation with respect to the other—show lines [21]. Indeed, in that case, in each time-frequency point (ω,n), one source clearly dominates the others, which implies that both observations are proportional to the dominant source and are thus linked through a linear relationship, hence the lines in the scatter plots. However, when the reverberation time increases, the disjointedness of the sources—i.e., the probability of source dominance in each time-frequency point—is less-likely to be satisfied [43] and the lines tend to be "blurry" or to disappear, as convolutive mixtures are frequency dependent.

As a consequence, analysing the scatter plots of the spectrograms of the CVR recordings might allow to analyse the nature of the source signals met in the observation signals. This phenomenon is illustrated on Figure 5, showing the scatter plots of $|X_1(\omega,n)|$ with respect to $|X_2(\omega,n)|$ for three kinds of signals, i.e., the pilot's voice and the radio (top), the pilot's voice and an announcement to passengers (middle), and the pilot's voice and an alarm (bottom). The plots have been computed using STFTs of short audio segments containing overlapping speech on Channels 1 and 2, i.e., $x_1(t)$ and $x_2(t)$ respectively, from 3 different CVR recordings. The top plot corresponds to the scenario depicted in Eq. (9)—which tends to show its validity—while the other plots correspond to more complex mixtures to separate.

Moreover, it should be noticed that the above models are defined for a fixed position. However, in practice, the aircraft pilots can turn their head, which might result in less selectivity to the cockpit sounds and in particular to the loudspeakers. Even if we did not model this phenomenon in this paper, we would like to emphasise the fact that many BSS methods have been extended to time-varying adaptive mixtures.

In addition, despite our efforts, we found that separating real CVR recordings was far more difficult than simulated ones using our model. This means that there remain some effects that we did not model and which need to be investigated. We particularly think that these effects are due to the aircraft system, prior to the recording in the CVR, e.g., dynamical filtering, pre-amplification, and gain control (with possibly clipping effects) of each pilot's microphone, the presence or the absence of anti-feedback filters. Lastly, we found one specific CVR model which quantises differently the signals with respect to their amplitude. This has some non-negligible effect on the BSS outputs.

To illustrate these different issues, we show on Fig. 6 the variability over speech segments selected from several CVR recordings where a mixture of one one pilot's voice and the radio channel has been found. According to Eq. (9), this mixture is LI and, as the sources are sparse in the TF domain, their scatter plots should consist of two lines. This is true when no additional effect is applied to the signals, as we can see on the left plot of Fig. 6. However, one may encounter situations when one channel is clipped (see the middle plot of Fig. 6) while—as already mentioned—we found one specific CVR model which quantises—with nonuniform quantisation—the CVR signals (right plot of Fig. 6). These specific issues have significant consequences on the source sparsity in the TF domain. However, they are out of the scope of this paper. In particular, deeply in-

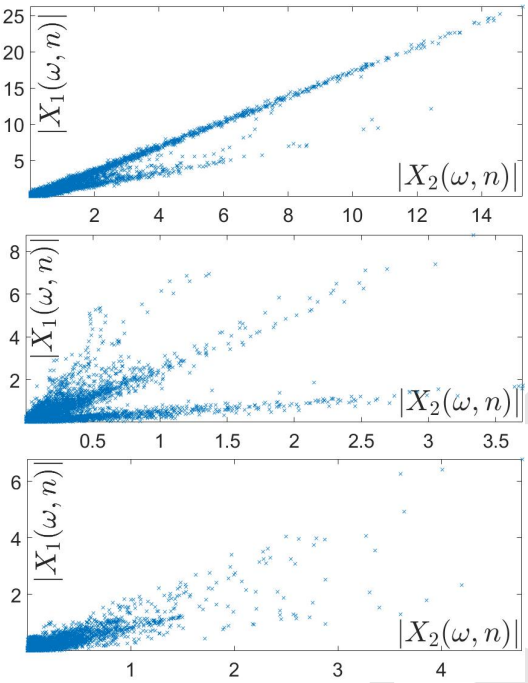

Fig. 5. Scatter plots for several source signals in real CVR recordings. From top to bottom: speech / radio signals; speech / passenger address; speech / cockpit alarm.

vestigating their effect on the BSS performance as well as proposing BSS methods which take them into account are let for future work.

Still, except for these specific situations, state-of-the-art BSS methods can be applied to CVR signals. This is the aim of the next section.

3 APPLICATION OF BSS METHODS TO CVR SIGNALS

3.1 Insights from a preliminary investigation

We do not have access to the true sources in real CVR recordings. This makes the evaluation of BSS methods quite difficult, as the classical objective performance criteria [44] cannot be used. As an alternative, we could firstly investigate their performance on simulations satisfying our proposed model. We then found that most of the tested BSS techniques were able to separate the sources in simulations while a similar scenario on real CVR recordings was much harder for all of them. This is probably due to the fact that, as mentioned above, the aircraft system may add some audio effects which are not taken into account in our model. This may explain why we found that, in SCA methods, estimating the sources using TF binary masking [5] was far better than after applying a sparsity-based technique [45], while the opposite result is usually found for "classical" audio mixtures. Indeed, binary masking is quite robust to some un-modeled effects (for example, it allows to separate convolutive mixtures of sources using a simpler model [5]). However, this comes at the price of generating more musical noise than alternative techniques such as [45].

3.2 Experimental protocol

As a result of the above insights, we propose a *subjective* experimental analysis which copes with typical BEA investigations. We limit the size of the corpus in this evaluation, so that it can be performed by BEA audio analysts. To that end, it consists of 25 anonymised speech segments for the scenario considering the superimposed speech of pilots and 21 speech segments for the scenario in which a pilot's voice is covered by the radio. These audio segments have a duration between 7 and 25 seconds and come from 15 distinct non-major incidents occurring during parking, taxi, takeoff, cruise, approach, and landing. They contain the voices of 12 men and 3 women, recorded by 4 types of CVRs.

We focused on the most widely installed CVR manufacturers (Honeywell 6022, Honeywell 6020, L3COM A100, and L3COM FA2100) and considered both magnetic-tape and solid-state memory CVR generations (12 solid-state CVR and 3 magnetic-tape CVR). These recordings also cover a set of 10 aircraft types from major aircraft manufacturers (Airbus A318, A319, A320, A321, A330, ATR-42, ATR-72, Boeing B777, Embraer EMB145, and Fokker FK100). While we consider BSS problems with only two sources, the separation process is not that easy to perform on real CVR signals, as we will see below.

We now investigate the potential benefits of using BSS outputs to help the audio analyst transcribe unintelligible superimposed speech signals. In this study, we decided not to investigate the performance of ICA/IVA approaches. Indeed, anticipating the use of BSS for BEA investigations in a near future, CVR signals may correspond to determined or under-determined mixtures while most ICA/IVA methods were proposed for (over-)determined mixtures only. Moreover, as already highlighted, the mixtures may evolve during time and the mixing parameters should be computed over small time intervals / STFT windows. However, the independence assumption of speech signals is questionable in that case [46]. Lastly, SCA methods outperform ICA ones [22, 23] and are more versatile, as they can process both the over-determined and the under-determined mixtures. Similarly, we decided not to investigate deep learning methods because they require training data, whose quality may have a significant impact on the BSS performance. Such a behaviour is not acceptable for regulatory investigations conducted by BEA. As a consequence, we aim to investigate the performance of SCA and NMF methods. We choose to use BSS methods whose sources are free and accessible.

We have chosen three BSS methods which all apply in the STFT representation domain, i.e., two SCA and one NMF methods. The first considered SCA method is DEMIX[2] [7]—used here in its version for LI mixtures—seeks to count the sources and to estimate the LI mix-

[2]We investigated the performance of much more methods in preliminary tests. We then found that DEMIX provided the same BSS enhancement as TIFROM [22] and TIFCORR [23], except in a few cases where it was outperforming them. This is the reason we only keep DEMIX in these tests. Moreover, we investigated

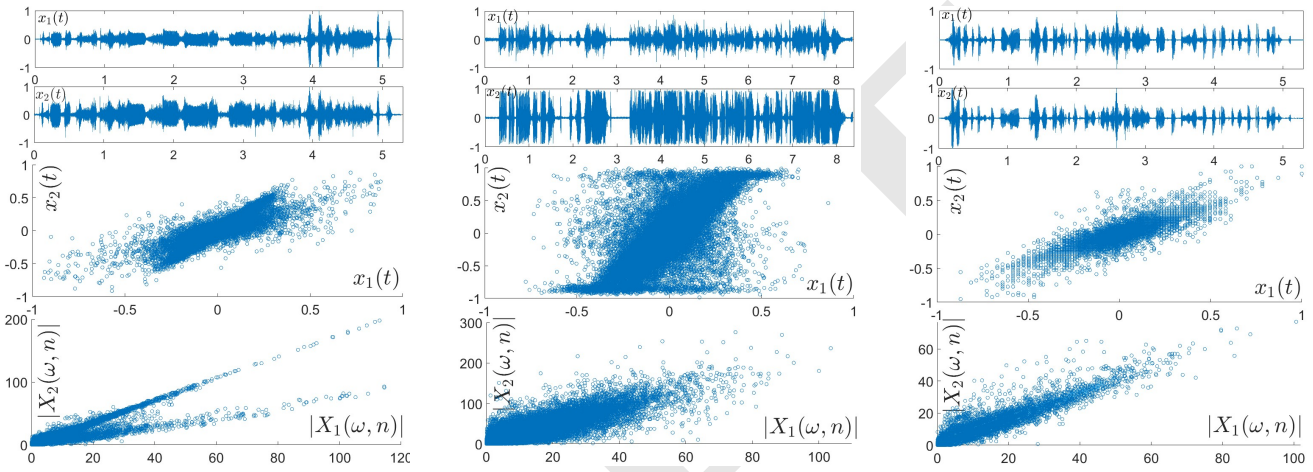

Fig. 6. Examples of CVR variability for several mixtures of one pilot's voice and the radio signals. Top: waveforms wrt. time. Middle: scatter plot of the observations in the time domain. Bottom: scatter plot of the observations in the TF domain. From left to right: CVR signals without visible quantisation effect nor clipping. Middle: Case of one clipped CVR channel. Right: Case of quantised CVR signals with non-uniform quantisation.

ing matrix using a dedicated clustering approach. The second method we investigate is a convolutive SCA technique named UCBSS [47] which estimates the mixing filters. For both SCA methods—and as explained in Subsect. 3.1—once the mixing parameters are estimated, the sources are estimated using binary masking [5]. Lastly, we also evaluate the enhancement provided by a multichannel NMF method [29]. In particular, we chose the expectation-maximisation algorithm and the sources are lastly estimated by applying a Wiener filter.

We now introduce the experiments done with the evaluation corpus. An audio analyst first produces a transcription of the segments of the evaluation corpus, indicating uncertain or unintelligible terms with a question mark *(?)*. At this stage the number of question marks is not necessarily reflecting the real amount of unintelligible words. The number of question marks is hence changed in the references before scoring depending on the number of words available in the output of all BSS methods. Fig. 7 shows two examples of such transcribed segments. The considered segments all contain at least one word that could not be transcribed with certainty by the analyst. A sequence of several unintelligible words are replaced by a sequence of question marks with the corresponding number of tokens. This primary transcription represents the best and the most honest result a CVR audio analyst can reach using his knowledge and skills.

The three BSS methods are then independently applied to Channels 1 and 2 over the corresponding segments and the audio analyst can use and listen to both BSS outputs as many times as he wants to replace the question marks with the correct words. Lastly, as we do not aim to evaluate the individual BSS performance, we consider a scenario where the audio analyst get all the BSS outputs—i.e., 6 outputs in this paper—to help him to transcribe CVR recordings.

- First example
  Pilot: première à gauche ouais c'est là on va on va se garer et puis après on (?)
  CTRL: three two zero push back approved
- Second example
  CDB: euh (?)
  FO: ça on avait (?)

Fig. 7. Examples of transcriptions without BSS.

3.3 Obtained results

To evaluate the experiment introduced in the previous subsection, we propose two performance criteria. The first one corresponds to the proportion of segments with improved transcription. Indeed, each tested BSS method is applied on each audio segment. When he listens to BSS outputs, if he is able to understand at least one non-transcribed word or sequence of words, without any doubt, the audio analyst considers that the BSS method succeeds in enhancing the signals. Once all the segments have been studied, we are then able to derive the proportion of improved segments for each method, as one may see on the top part of Table 1.

Then, we only consider the above improved segments and the audio analyst analyses again the CVR recordings while using the corresponding BSS outputs in addition. The performance of each method is evaluated in terms of Unintelligible Word Recognition Rate (UWRR), which is the percentage of initially unintelligible words that could ultimately be transcribed after application of BSS. In particular, we first estimate the number of initially unintelligible worlds by summing the number of interrogative points. Then, after applying BSS, when an unintelligible segment can be transcribed, we can count the number of transcribed words and we update the number of initially unintelligible

the performance of a silence-based technique [26]—i.e., an SCA method with a relaxed source sparsity assumption—but it provided a poor enhancement in all our tests. Lastly, we also tried a method designed for anechoic mixtures—i.e., DUET [5]—but it almost always provided a lower BSS enhancement than DEMIX.

ones. This tends to possibly overestimate the UWRR values, as excerpts which remain unintelligible are counted as a single word. Still, in absence of data with ground truth, we keep this criterion which should be seen as an optimistic measure. The obtained UWRRs are shown on the bottom part of Table 1.

Let us now focus on the segment transcription rates. First, one may notice that the "pilot / pilot" scenario seems harder to improve, as the better-suited method can improve 28% of the segments. On the contrary, the "pilot / radio" scenario seems a bit simpler, as the worse-suited method can improve the transcription of 28.5% of the segments. Moreover, in all the scenarios, UCBSS provides the lowest enhancement. In addition, NMF and DEMIX are the best approaches in terms of "pilot / radio" and "pilot / pilot" segment transcription improvement (with a proportion of 38% and 28%, respectively), respectively. Lastly, it is very interesting to notice that when we allow the audio analyst to listen to the outputs of the 3 methods, the transcription of 44% and 66% of the "pilot / pilot" and "pilot / radio" segments is improved, respectively. This means that the tested BSS methods do not enhance the same segments.

We now focus on UWRR. Let us recall that the proportions are derived from the transcribed words in the improved segments only. First of all, these proportions for "pilot / pilot" segments are quite high, as 50 to 57.5% of unintelligible words can be transcribed by using one of the tested BSS methods. This proportion is much higher for "pilot / radio" segments as it ranges between 56 and 70%. Still, from a BEA investigator point of view, the most interesting result we get is due to the combination of all the BSS outputs. In that case, the above rates are equal to 80 and 89.6% for "pilot / pilot" and "pilot / radio" segments, respectively. Lastly, it should be noticed that the tested multichannel NMF method not only provides the highest number of transcribed words which were initially unintelligible but also produces the most intelligible sources on a majority of audio segments.

To illustrate the ability of BSS to help the BEA audio analyst in his transcription tasks, Fig. 8 provides the same example as in Fig. 7, except that BSS outputs were here used for the transcription.

4 CONCLUSION ET PERSPECTIVES

In this paper, we investigated the enhancement which may be provided by blind source separation methods on real recordings of some cockpit voice recorders, more commonly called "black boxes". To that end, we firstly proposed a CVR mixing model, obtained by reverse engineering. While we highlighted the limits of this model, we could apply classical BSS methods on a corpus of real CVR recordings. We found that combining the ouptuts of various BSS methods can be really helpful for BEA audio analysts, with a word recognition rate of initially unintelligible words ranging from 80 to 90%. Still, while the obtained results are promising, there remain challenges.

First of all, some of the audio effects due to the aircraft system were not taken into account, e.g., the clipping effects. Combining BSS and declipping has been investigated and it would be interesting to measure any improvement by extending the tested multichannel-NMF to that case. Moreover, the pilots may move during the flight, which may result in time-varying mixing parameters. This should also be considered in future investigations.

Lastly, we would like to emphasise that other CVR channels are available. In particular, the CAM was not considered in this paper as it is not sampled at the same frequency rate than the pilots' microphones. However, it provides some interesting information, e.g., the aircraft engine noise, or the announcements to passengers, which are hardly recorded by the pilots' microphones[3]. Jointly separating the pilots' channels and the CAM will be investigated in the future.

- First example
  Pilot: première à gauche ouais c'est là on va on va se garer et puis après on fera le reste hein
  CTRL: three two zero push back approved
- Second example
  CDB: euh l'APU est démarré
  FO: ça on avait prévenu

Fig. 8. Examples of transcriptions using BSS outputs.

5 ACKNOWLEDGMENT

This work has been funded in part by BEA.

6 REFERENCES

[1] P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (Elsevier, 2010), doi: 10.1016/C2009-0-19334-0.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979 (1953 Sept.).

[3] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," presented at the *International Conference on Independent Component Analysis and Signal Separation*, pp. 734–741 (2009), doi:10.1007/978-3-642-00599-2_92.

[4] A. Yeredor, "Time-delay estimation in mixtures," presented at the *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003*, vol. V, pp. 237–240 (2003), doi:10.1109/ICASSP.2003.1199912.

[5] 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 July), doi:10.1109/TSP.2004.828896.

[3] However, please note that the CVR channel corresponding to the third pilot may be used to store the announcements to the passengers.

		DEMIX	UCBSS	NMF	3 methods
Segment transcription improvement rate	pilot / pilot	28%	16%	20%	44%
	pilot / radio	33%	28.5%	38%	66%
UWRR	pilot / pilot	50%	55%	57.5%	80%
	pilot / radio	56%	63.6%	70%	89.6%

Table 1. Reached performance on CVR transcription after applying BSS methods.

[6] M. Puigt and Y. Deville, "Time–frequency ratio-based blind separation methods for attenuated and time-delayed sources," *Mechanical Systems and Signal Processing*, vol. 19, no. 6, pp. 1348–1379 (2005 Nov.), doi:10.1016/j.ymssp.2005.08.003.

[7] S. Arberet, R. Gribonval, and F. Bimbot, "A Robust Method to Count and Locate Audio Sources in a Multichannel Underdetermined Mixture," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121–133 (2010 Jan.), doi:10.1109/TSP.2009.2030854.

[8] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960 (2012 Aug.), doi:10.1016/j.sigpro.2011.09.032.

[9] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2193–2206 (2013 Oct.), doi:10.1109/TASL.2013.2272524.

[10] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2807–2820 (1999 Oct.), doi:10.1109/78.790661.

[11] M. Puigt, A. Griffin, and A. Mouchtaris, "Post-nonlinear speech mixture identification using single-source temporal zones & curve clustering," presented at the *2011 19th European Signal Processing Conference*, pp. 1844–1848 (2011).

[12] M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Blind separating convolutive post-nonlinear mixtures," presented at the *3rd Workshop on Independent Component Analysis and Signal Separation (ICA 2001)*, pp. 138–143 (2001).

[13] Ç. Bilen, A. Ozerov, and P. Pérez, "Joint audio inpainting and source separation," presented at the *12th International Conference on Latent Variable Analysis and Signal Separation*, pp. 251–258 (2015), doi:10.1007/978-3-319-22482-4_29.

[14] M. Puigt, E. Vincent, Y. Deville, A. Griffin, and A. Mouchtaris, "Effects of audio coding on ICA performance: An experimental study," presented at the *2013 IEEE 11th International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics*, pp. 1–6 (2013), doi:10.1109/ECMSM.2013.6648949.

[15] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1699–1712 (2013 Aug.), doi:10.1109/TASL.2013.2260153.

[16] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314 (1994), doi:10.1016/0165-1684(94)90029-9.

[17] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," presented at the *6th International Conference on Independent Component Analysis and Signal Separation (ICA 2006)*, pp. 165–172 (2006), doi:10.1007/11679363_21.

[18] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871 (2007 Aug.), doi:10.1016/j.sigpro.2007.01.010.

[19] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. Signal Process.*, vol. 46, no. 11, pp. 2888–2897 (1998 Nov.), doi:10.1109/78.726803.

[20] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 386–389 (2004 March), doi:10.1109/LSP.2003.819343.

[21] R. Gribonval and M. Zibulevsky, "Chapter 10 - Sparse component analysis," in P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation*, pp. 367–420 (Academic Press, Oxford, 2010), doi:10.1016/B978-0-12-374726-6.00015-1.

[22] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403 (2005 July), doi:10.1016/j.sigpro.2005.02.010.

[23] Y. Deville and M. Puigt, "Temporal and time-frequency correlation-based blind source separation methods. Part I: Determined and underdetermined linear instantaneous mixtures," *Signal Processing*, vol. 87, no. 3, pp. 374–407 (2007 March), doi:10.1016/j.sigpro.2006.05.012.

[24] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 992–996 (2005).

[25] F. M. Naini, G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "Estimating the mixing matrix in Sparse Component Analysis (SCA) based on partial k-dimensional subspace clustering," *Neurocomputing*, vol. 71, no. 10-12, pp. 2330–2343 (2008 June), doi:10.1016/j.neucom.2007.07.035.

[26] B. Rivet, "Blind non-stationnary sources separation by sparsity in a linear instantaneous mixture," presented at the *International Conference on Independent Component*

Analysis and Signal Separation, pp. 314–321 (2009), doi: 10.1007/978-3-642-00599-2_40.

[27] A. Cichocki, R. Zdunek, and S.-i. Amari, "Nonnegative matrix and tensor factorization [lecture notes]," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 142–145 (2007), doi:10.1109/MSP.2008.4408452.

[28] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," presented at the *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177 – 180 (2003 11), doi:10.1109/ASPAA.2003.1285860.

[29] A. Ozerov and C. Févotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563 (2010 March), doi:10.1109/TASL.2009.2031510.

[30] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," presented at the *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70 (2015), doi:10.1109/ICASSP.2015.7177933.

[31] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," presented at the *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3437–3440 (2009), doi: 10.1109/ICASSP.2009.4960364.

[32] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207 (2010 Aug.), doi: 10.1109/TASL.2009.2031694.

[33] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," presented at the *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 257–260 (2011), doi:10.1109/ICASSP.2011.5946389.

[34] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320 (2015 May), doi:10.1109/TSP.2015.2404577.

[35] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664 (2016 Sept.), doi: 10.1109/TASLP.2016.2580946.

[36] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254* (2021), doi: 10.48550/arXiv.1911.13254.

[37] T. Li, J. Chen, H. Hou, and M. Li, "Sams-net: A sliced attention-based neural network for music source separation," presented at the *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5 (2021), doi:10.1109/ISCSLP49672.2021.9362081.

[38] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726 (2018 Oct.), doi: 10.1109/TASLP.2018.2842159.

[39] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219 (2019 May), doi:10.1109/JSTSP.2019.2908700.

[40] F.-B. Hsiao, S.-Y. Han, S.-C. Hsieh, and L. K. Wang, "Sound source separation and identification for aircraft cockpit voice recorder," *Journal of Aerospace Computing, Information, and Communication*, vol. 1, no. 12, pp. 466–483 (2004 Dec.), doi:10.2514/1.11266.

[41] D. Nie, X. Li, and G. Qiao, "Cockpit noise enhancement for aircraft type recognition in short-wave speech communication," presented at the *International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'13)*, p. 1 (2013).

[42] A. Martinez, L. Sanchez, and I. Couso, "Interval-valued Blind Source Separation Applied to AI-based Prognostic Fault Detection of Aircraft Engines," *Journal of Multiple-Valued Logic & Soft Computing*, vol. 22, p. 151 (2014).

[43] S. Schulz and T. Herfet, "On the window-disjoint-orthogonality of speech sources in reverberant humanoid scenarios," presented at the *11th International Conference on Digital Audio Effects DAFx–08*, pp. 241–248 (2008).

[44] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469 (2006 July), doi:10.1109/TSA.2005.858005.

[45] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," presented at the *7th International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, pp. 430–437 (2007), doi:10.1007/978-3-540-74494-8_54.

[46] M. Puigt, E. Vincent, and Y. Deville, "Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources," presented at the *8th International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, pp. 613–620 (2009), doi:10.1007/978-3-642-00599-2_77.

[47] 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 Jan.), doi: 10.1109/TASL.2009.2024380.

THE AUTHORS

Matthieu Puigt

Benjamin Bigot

Hélène Devulder

Matthieu Puigt graduated in 2003 from University Paul Sabatier Toulouse 3 (Toulouse, France) with an M.Sc. in signal, image processing, and acoustics. He received a Ph.D. in signal processing from the University of Toulouse (Toulouse, France) in 2007. He was Lecturer at the University Paul Sabatier Toulouse 3 from September 2007 to August 2009. He was Assistant Professor at the University for Information Science and Technology, in Ohrid, North Macedonia, from September 2009 to June 2010. From August 2010 to July 2012, he was a Marie Curie Postdoctoral fellow with the Signal Processing Lab, Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Crete, Greece. From September 2012 to August 2024, he was Associate Professor at the University of Littoral Cote d'Opale, in Calais and Saint-Omer, France while he is currently Professor at the same university. His research interests include sparse and nonnegative latent variable analysis methods and their applications to acoustics, environment monitoring, and hyperspectral imaging.

●

Benjamin Bigot received a Master degree in Signal processing from the University Paul Sabatier Toulouse 3 in 2007, and a Ph.D. degree in Computer Science with application to Automatic Speech Processing from the University Paul Sabatier Toulouse 3 in 2011. He is currently a senior air accident safety investigator, specialised in the audio analysis of Cockpit Voice Recorders at the French Air Accident Investigation Bureau (BEA for Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile). Aside his safety investigator duty, he is empowered in leading research and development projects in audio and speech processing for the specific needs of audio analysts of BEA's Engineering Department.

●

Hélène Devulder graduated from Grenoble INP – ENSE3 in 2020. In 2019, she was also a visiting student at Pohang University of Science and Technology, South Korea. After she graduated, she changed her career plans and she is now teacher in a primary school in Northern France.