A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture

Simon Arberet, Rémi Gribonval, Frédéric Bimbot

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture

Simon Arberet — Rémi Gribonval — Frédéric Bimbot

N° 6593

July 2008

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6593--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture

Simon Arberet*, Rémi Gribonval†, Frédéric Bimbot*

Thème COG — Systèmes cognitifs
Équipes-Projets METISS

Rapport de recherche n° 6593 — July 2008 — 29 pages

Abstract: We propose a method to count and estimate the mixing directions and the sources in an underdetermined multichannel mixture. Like DUET-type methods, the approach is based on the hypothesis that the sources have time-frequency representations with limited overlap. However, instead of assuming essentially disjoint representations, we only assume that, in the neighbourhood of *some* time-frequency points, only one source contributes to the mixture: such time-frequency points can provide robust local estimates of the corresponding source direction. At the core of our contribution is a local confidence measure –inspired by the work of Deville on TIFROM– which detect the time-frequency regions where such a robust information is available. A clustering algorithm called DEMIX is proposed to merge the information from all time-frequency regions according to their confidence level. Two variants are proposed to treat instantaneous and anechoic mixtures. In the latter case, to overcome the intrinsic ambiguities of phase unwrapping as met with DUET, we propose a technique similar to GCC-PHAT to estimate time-delay parameters from phase differences between time-frequency representations of different channels. The resulting method is shown to be robust in conditions where all DUET-like comparable methods fail: a) when time-delays largely exceed one sample; b) when the source directions are very close. As an example, experiments show that, in more than **65**% of the tested stereophonic mixtures of six speech sources, DEMIX-Anechoic correctly estimates the number of sources and outperforms DUET in the accuracy, providing a distance error 10 times lower.

Key-words: blind source separation, multichannel audio, direction of arrival, delay estimation, sparse component analysis

* CNRS
† INRIA

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Une méthode robuste pour compter, localiser et séparer les sources audio d'un mélange multicanal sous-déterminé

Résumé : Nous proposons une méthode pour compter et estimer les directions et les sources d'un mélange multicanal sous-déterminé. De la même façon que pour les méthodes de type DUET, l'approche est basée sur l'hypothèse que les sources ont des représentations temps-fréquence qui se chevauchent peu. Cependant, plutôt que de supposer que les représentations aient des supports disjoints, nous supposons seulement que, dans le voisinage de *quelques* points temps-fréquence, seulement une source contribue au mélange: de tels points temps-fréquence peuvent fournir des estimations locales robustes des directions des sources correspondantes. Une de nos contributions majeures est une mesure de confiance locale –inspirée des travaux de Deville sur TIFROM– qui détecte les régions temps-fréquence où de telles informations sont disponibles. Nous proposons un algorithme de clustering appelé DEMIX qui permet de traiter l'information de toutes les régions temps-fréquence suivant leur niveau de confiance. Deux variantes de l'algorithme sont proposées afin de traiter le cas instantané et le cas anéchoïque. Dans ce dernier cas, afin de résoudre le problème intrinsèque de repliement de phase rencontré dans DUET, nous proposons une technique proche de GCC-PHAT pour estimer les paramètres de délai à partir des différences de phase entre les représentations temps-fréquence des différents canaux. La méthode résultante se montre robuste dans les situations où les méthodes de type DUET échouent: a) quand les délais sont très supérieurs à un échantillon; b) quand les directions des sources sont proches. Les expériences montrent que pour un mélange stéréophonique constitué de six sources, DEMIX-Anechoic estime correctement le nombre de sources dans plus de **65**% des cas, et obtient des estimations des directions avec une erreur moyenne plus de 10 fois inférieure à celle de DUET.

Mots-clés : separation de source aveugle, audio multicanal, direction d'arrivée, estimation de délais, analyse en composantes parcimonieuses

1 Introduction

In many situations like medical imaging, musical or meeting recording, the observed data is a measurement of several signals which have been mixed together, and it is sometime very usefull to know what the original signals (called sources) were. In the context of audio sources, the measured signals are often on only two channels, that is the well known stereophonic case, and the number of sources are often higher than the number of channels.

In this article, we consider the problem of separating several audio sources from two or more mixtures when there may be more sources than available mixtures, with an emphasis on stereophonic audio mixtures. Our approach relies on the now classical time-frequency masking framework [1, 2] with a two-step approach: a first step consists in estimating the number of sources and their mixing directions; in a second step the sources are separated using appropriate adaptive time-frequency masks.

Our main contribution is a new technique to perform the first step, called DEMIX (Direction Estimation of Mixing matrIX) [3, 4] which relies on the weak assumption that in *some* time-frequency regions, essentially one source contributes to the mixture. In such regions, the intensity difference and phase-difference between channels provide information on the direction of the corresponding source. The proposed technique estimates both the number of sources and their mixing directions through a new clustering algorithm. This clustering algorithm gives more weight to more reliable time-frequency regions, according to a local confidence measure similar to the one proposed in TIFROM [5], [6].

We demonstrate with extensive experimental studies the ability of our approach a) to blindly estimate the number of sources in anechoic mixtures of up 6 sources; b) to robustly estimate time-delays than can largely exceed one sample, thanks to the use of a technique similar to GCC-PHAT [7], unlike DUET [2] which is essentially limited to delays below one sample; c) to outperform DUET in the accuracy of the estimation of direction by a distance error at least 10 times lower; d) to robustly estimate nearby source direction with a constant relative precision better than 10^{-3}. In addition to DEMIX, we propose and demonstrate a variant of time-frequency masking to perform the separation step. The variant, which coincides with standard time-frequency masking for anechoic mixtures with delays below one sample, has significantly better performance even with time-delays of several tens of samples.

1.1 Anechoic mixture model and source separation

The mixture of N audio sources on M channels can be formulated by the anechoic mixture model :

$$x_m(\mathrm{t}) = \sum_{n=1}^{N} a_{mn} s_n(\mathrm{t} - \delta_{mn}) + n_m(\mathrm{t}), \qquad 1 \leq m \leq M \tag{1}$$

In this model, each source contributes to each microphone only through the direct acoustic path, that is to say with no reflection on walls or obstacles. The parameters $a_{mn} \in \mathbb{R}$ represent the gain (or the attenuation) and δ_{mn} the time-delay corresponding to the path between the n-th source and the m-th microphone. The problem we address in this paper is the estimation of the

number of sources N, the mixture parameters a_{mn} and δ_{mn}, and the source signals $s_n(\mathrm{t})$ from the only observation of the noisy mixture signals $x_m(\mathrm{t})$.

When the number N of sources is known, this is called a blind source separation (BSS) problem [8]. If in addition the number M of sensors does not exceed the number of sources, if the problem is noiseless, and if the delays equal zero, this is an (over)determined linear instantaneous BSS problem, which is known to admit a unique solution (up to gain, sign, permutation and shift indeterminacies) under the mild assumption that the sources are statistically independent and non-Gaussian. Here, the number of sources is unknown and might exceed the number of sensors, yielding an underdetermined BSS problem. Indeed, when dealing with an unknown stereophonic musical recording (with $M = 2$ channels), it is only realistic to assume that the number of instruments that are playing together exceeds two and is not known in advance.

Without loss of generality, we assume that $\sum_{m=1}^{M} a_{mn}^2 = 1$ and that $\delta_{1n} = 0$ and $a_{1n} \geq 0$ for $1 \leq n \leq N$, which means that we fix the gain, sign and shift indeterminacies of the problem. If in addition $\delta_{mn} = 0$ for all m and n the mixture is indeed instantaneous. Taking the Short Time Fourier Transform (STFT) $X_m(t, f)$ of each channel $x_m(\mathrm{t})$ of the mixture, the mixing model is approximately written in complex matrix form in the time-frequency domain as $\mathbf{X}(t, f) = \mathbf{A}(f)\mathbf{S}(t, f) + \mathbf{N}(t, f)$ for each time frame t and normalized frequency $0 \leq f \leq 1/2$, where bold letters such as $\mathbf{X}(t, f)$ or $\mathbf{S}(t, f)$ denote column vectors $[X_1(t, f), \ldots, X_M(t, f)]^T$ or $[S_1(t, f), \ldots, S_N(t, f)]^T$, and $\mathbf{A}(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_N(f)]$ is the $M \times N$ mixing matrix which columns $\mathbf{a}_n(f) = [a_{1n}, a_{2n}e^{-2i\pi\delta_{2n}f}, \ldots, a_{Mn}e^{-2i\pi\delta_{Mn}f}]^T$ are related to the source locations.

In the stereophonic case ($M = 2$), each column of $\mathbf{A}(f)$ can be written as a two-dimensional vector :

$$\mathbf{a}_n(f) = \begin{bmatrix} \cos\theta_n \\ \sin\theta_n \cdot e^{-2i\pi\delta_n f} \end{bmatrix} \in \mathbb{C}^2. \tag{2}$$

The parameter $\theta_n \in (-\pi/2, \pi/2]$ characterizes the *intensity difference* (ID) between channels, and a possible sign difference if $\theta_n < 0$; the parameter $\delta_n \in \mathbb{R}$ characterizes the *time delay* between channels. We will generally refer to the pair (θ_n, δ_n) as the (mixing) direction of the n-th source. For the case $M > 2$ channels, we can generalize this pair by splitting the direction of the n-th source into its *intensity* defined by $\mathrm{abs}\,(\mathbf{a}_n(f)) \in \mathbb{R}^M$ with $\|\,\mathrm{abs}(\mathbf{a}_n(f))\|^2 = 1$, and delays $\Delta_n = [\delta_{1n}, \delta_{2n}, \ldots, \delta_{Mn}]^T \in \mathbb{R}^M$ with $\delta_{1n} = 0$.

1.2 Related work about the estimation of the mixing directions

Several existing methods attempt to estimate the mixing directions of the sources from a time-frequency representation $\mathbf{X}(t, f)$ of the mixture. DUET-type methods [2, 1] rely on the assumption that the mixed sources have essentially disjoint time-frequency supports, that is to say in *most* time-frequency points, only one source has a non-negligible contribution. This is related to the sparsity assumption on the time-frequency representation of the sources. TIFROM [5] exploits the weaker assumption that for each source, there is *at least one* time-frequency region where this source is dominant. Our approach relies on the same assumption as TIFROM. This means that it can still estimate the directions if in most

of the time-frequency plane, several (or even all) sources are similarly active, provided that for each source one can still find a (tiny) time-frequency region where it emerges from the background. In the latter situation, it will however be difficult to go beyond the direction estimation step, and separating the sources might actually be a daunting task.

When at most one source actively contributes to a time-frequency point (t, f), there is an index $1 \leq n(t, f) \leq N$ such that $|S_{n(t,f)}(t, f)| \gg |S_n(t, f)|$, $n \neq n(t, f)$, so the mixing model indicates that $\mathbf{X}(t, f) \approx \mathbf{A}(f)\mathbf{S}(t, f) \approx S_{n(t,f)}(t, f) \cdot \mathbf{a}_{n(t,f)}(f)$ and the ratio $R_{21}(t, f) := X_2(t, f)/X_1(t, f)$ satisfies :

$$R_{21}(t, f) \approx \tan \theta_{n(t,f)} \cdot e^{-2i\pi\delta_{n(t,f)}f}.$$

So if the sources have disjoint time-frequency supports, then all data points $\mathbf{X}(t, f)$ will be aligned along the directions $\mathbf{a}_{n(t,f)}(f)$. Also, if the sources are sparse, the data points $\mathbf{X}(t, f)$ show a clear tendency to *cluster* along $\mathbf{a}_{n(t,f)}(f)$ [1]. This can be seen on the scatter plot of points $\mathbf{X}(t, f)$, which is a simple tool we will use latter in this paper (see for example figures 1 and 2). A common approach to estimate the mixing directions is thus based on an clustering algorihm applied on the points of the scatter plots.

In DUET [2] the ratios $R_{21}(t, f)$ are computed for each time-frequency point and used to compute a local estimate of the intensity difference $\theta(t, f) := \tan^{-1} |R_{21}(t, f)|$ and the delay $\delta(t, f) := -\frac{1}{2\pi f} \angle R_{21}(t, f)$ where $\angle z \in (\pi, \pi]$ is the principal argument of a complex number z. This approach is perfectly valid if the true delay is below one sample and the gains a_{mn} are all positive, but it may fail otherwise because of phase unwrapping ambiguities. In a sense, the problem is that a single time-frequency point does not carry alone enough information to recover the corresponding delay $\delta_{n(t,f)}$. To recover it, it is therefore necessary to gather the information coming from several time-frequency points where the same source is active at different frequencies, which raises two issues: 1) How to find several points associated to the same source ? 2) How to efficiently deduce the delay, with no ambiguity, given several points associated to the same source ? A new approach to solve these issues is presented in Section 4. The proposed technique is able to estimate *time delays which can largely exceed one sample*, as illustrated in Section 5.3.

1.3 Local confidence measure

The TIFROM assumption is more realistic than the DUET one, because in many audio mixture, there can be a majority of time-frequency points where several sources are simultaneously active. It is clear that time-frequency points where several sources are equally active yield local estimates completely unrelated to the true directions of any source. The problem is thus how to detect time-frequency points where only one source is essentially active. Inspired by TIFROM [5] and related work [6] we describe in Section 2 how to compute a *local confidence measure* $\mathcal{T}(t, f)$ which estimates how likely it is that a single source is active around the time-frequency point (t, f).

1.4 Clustering algorithms for unknown number of sources

When it comes to actually clustering local estimates of the source directions to get a global estimate of the number of sources and their directions, many au-

thors have chosen to use a weighted smoothed histogram [2], where the amount of smoothing is determined by the shape of a "potential function" [1]. One of the difficulties with this approach consists in ajusting how much smoothing must be performed on the weighted histogram to resolve close directions without introducing spurious peaks. Moreover, the choice of the weights is also of importance. The classical approach, which consists in giving more weight to local estimates if they are associated to a time-frequency point with more energy, might prevent the clustering step from properly discovering the direction of a source of weak energy. Instead of using a fixed potential function and weights based on the local energy, we introduce in Section 3 a new clustering algorithm which relies on the local confidence measure introduced in Section 2 and a statistical model described in the Appendix A. Experiments reported in Section 5 illustrate its ability to adaptively find the number of sources and their directions. An important feature of the proposed clustering algorithm is that its accuracy does not depend on a prior choice of a smoothing parameter.

1.5 Time-frequency masking of mixtures with large delays

Estimating the source directions is only the first step of a source separation process. In order to get estimates of the sources, the second step is often based on time-frequency masking [2], followed by an inverse STFT (by the overlap-add method). For each source, a time-frequency mask is built which indicates the time-frequency locations where it is considered as active, and the STFT of the mixture is multiplied by this mask before being inverted. The mask is built by correlating the estimated source directions with the mixture.

Even though this approach has been shown to be quite successful on anechoic mixtures with short delays between sources, it can completely fail if the delays become longer, as shown experimentally in Section 5.

We propose in Appendix B an alternate strategy which consists in replacing standard time-frequency masking by a variant where the correlation between the mixture and the direction of a source is computed after resynchronizing the channels according to the estimated time-delay between channels for the given source. Numerical results in Section 5 show the improvement over standard time-frequency masking obtained with this approach.

2 Principle of the approach

The approach we propose to estimate the mixing directions in the instantaneous case rely on the same assumption as TIFROM, that is for each source there is at least one time-frequency region where it is the only "visible" source. The first step of our method is a feature extraction step which can easily dicriminate time-frequency regions where essentially one source is active, from time-frequency regions where zero or more than one source is active. The second step of our method is the clustering algorithm which is defined in section 3.

2.1 Feature extraction

For each time-frequency region $\Omega_{t,f}$ "in the neighborhood" of the time-frequency point (t, f), the principle is to estimate two values:

1. the direction $\hat{\mathbf{u}}(\Omega_{t,f})$ of the most dominant source;

2. a *local confidence measure*, denoted $\widehat{\mathcal{T}}(\Omega_{t,f})$, which gets larger when the scatter plot of vectors $\mathbf{X}(\tau,\omega)$ in the region $\Omega_{t,f}$ points more strongly in the direction $\hat{\mathbf{u}}(\Omega_{t,f})$, that is when essentially one source is active in this region.

As the confidence measure $\widehat{\mathcal{T}}(\Omega_{t,f})$ can discriminate the cases where essentially one source is active from the other ones, it also discriminate cases where the direction $\hat{\mathbf{u}}(\Omega_{t,f})$ correspond to a true direction from the cases where $\hat{\mathbf{u}}(\Omega_{t,f})$ has few chance to point in one of the true directions.

To estimate the directions $\hat{\mathbf{u}}(\Omega_{t,f})$ and their corresponding local confidence measure, one can simply rely on Principal Component Analysis (PCA) and define $\hat{\mathbf{u}}(\Omega_{t,f})$ as the principal direction of the local scatter plot of vectors $\mathbf{X}(\tau,\omega)$ in the region $\Omega_{t,f}$, and $\widehat{\mathcal{T}}(\Omega_{t,f})$ a measure (defined in section 2.4) of how strongly it points in its principal direction.

2.2 Time-frequency regions

We consider two kinds of time-frequency regions around each time-frequency point (t,f) : a temporal neighborhood $\Omega^T_{t,f}$ and a frequency neighborhood $\Omega^F_{t,f}$. A discrete STFT with a window of size L computed with half overlapping windows and no zero-padding provides STFT values $\mathbf{X}(t,f)$ on the discrete time-frequency grid $t = kL/2$, $k \in \mathbb{Z}$ and $f = l/L$, $0 \leq l \leq L/2$. The time (respectively frequency) neighborhood of a time-frequency point (t,f) are defined by:

$$\Omega^T_{t,f} = \{(t + kL/2, f) \quad | \quad |k| \leq K\} \tag{3}$$

$$\Omega^F_{t,f} = \{(t, f + k/L) \quad | \quad |k| \leq K\}. \tag{4}$$

2.3 Real-valued and complex-valued local scatter plots

Each region Ω provides a complex-valued local scatter plot $\mathbf{X}(\Omega)$. It is a $M \times (2K+1)$ matrix with entries $\mathbf{X}(\tau,\omega)$, $(\tau,\omega) \in \Omega$ which will be used for anechoic mixtures. For linear instantaneous mixtures, since the directions $\mathbf{a}_n(f)$ of the sources are real-valued, a real-valued local scatter plot will be used instead. It corresponds to a $M \times (4K+2)$ matrix denoted $\mathbf{X}^{\mathbb{R}}(\Omega)$ with entries $\Re\mathbf{X}(\tau,\omega)$ and $\Im\mathbf{X}(\tau,\omega)$, $(\tau,\omega) \in \Omega$.

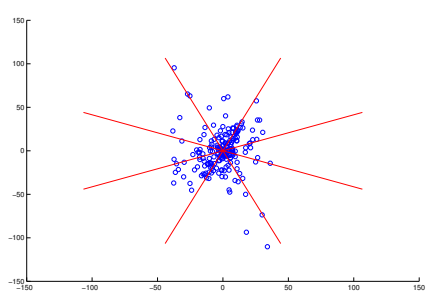
2.4 Principal Component Analysis and confidence measure

Performing a Principal Component Analysis (PCA) on the local scatter plot $\mathbf{X}(\Omega)$ (resp. $\mathbf{X}^{\mathbb{R}}(\Omega)$) we obtain a principal direction as a unit vector $\hat{\mathbf{u}}(\Omega) \in \mathbb{C}^M$ (resp. $\hat{\mathbf{u}}(\Omega) \in \mathbb{R}^M$) as well as the real-valued positive eigenvalues in decreasing order $\hat{\lambda}_1(\Omega) \geq \ldots \geq \hat{\lambda}_M(\Omega) \geq 0$ of the $M \times M$ complex Hermitian positive definite matrix $\mathbf{X}(\Omega)\mathbf{X}^H(\Omega)$ (resp. the real symmetric positive definite matrix $\mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$). We define the (empirical) confidence measure

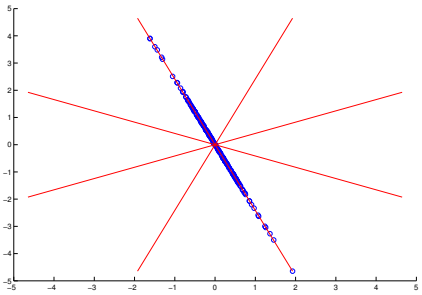
$$\widehat{\mathcal{T}}(\Omega) := \widehat{\lambda}_1(\Omega) \Big/ \frac{1}{M-1}\sum_{m=2}^{M} \widehat{\lambda}_m(\Omega)\ . \tag{5}$$

We will discuss in the Appendix why this measure can also be viewed as a signal to noise ratio between the dominant source and the contribution of the other ones plus noise. We will often express it in deciBel (dB) scale : $20\log_{10}(\widehat{\mathcal{T}}(\Omega))$.

Figure 1 shows the local scatter plot of $\mathbf{X}^{\mathbb{R}}(\Omega)$ in two time-frequency regions : one where many sources are simultaneously active, and another one where essentially one source is active. As expected by the theoretical results of the Appendix A, the confidence measure is high when essentially one source is active, and low when many sources are simultaneously active.



(a) Region where multiple sources contribute to the mixture. The confidence value is low (9.4 dB)



(b) Region where essentially one source contributes to the mixture. The confidence value is high (101.4 dB)

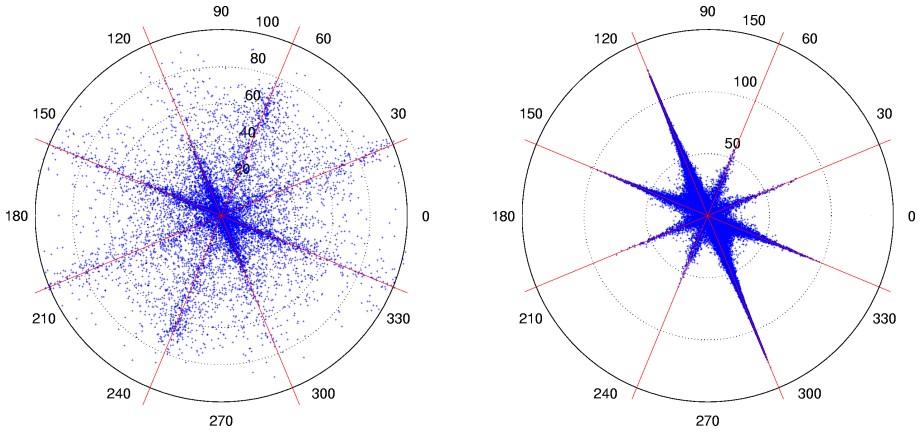
Figure 1: Local scatter plots in two time-frequency regions. Lines indicate true source directions. STFT window size is $L = 4096$.

Figure 2(a) displays the real-valued global scatter plot for all time-frequency points weighted by their energy, which is used in standard approaches to determine the source directions. In contrast, Figure 2(b) displays the collection of vectors $\pm 20\log_{10}\widehat{\mathcal{T}}(\Omega_{t,f}) \cdot \hat{\mathbf{u}}(\Omega_{t,f})$ obtained by PCA for all time-frequency regions of the signal. On both figures four lines indicate the angles corresponding to the true underlying directions. One can observe that points of figure 2(b) are better concentrate along the mixing directions than the one of figure 2(a), and thus, points of figure 2(b) should be a better candidate to estimate the mixing direction by a clustering algorithm, than the one of the standard approach. This will be confirmed experimentally.

3 Directions estimation Algorithms

In this section we describe the two proposed DEMIX (Direction Estimation of Mixing matrIX) algorithms. First we present DEMIX-Instantaneous, which is designed to estimate the directions of instantaneous mixtures, and second we present DEMIX-Anechoic, which is designed to estimate the directions of anechoic mixtures.

DEMIX algorithms belong to the categories of the sequential clustering algorithms, and are related to the Basic Sequential Algorithmic Scheme (BSAS) algorithm [9]. In sequential clustering algorithms, points are presented to the algorithm in a certain order. The basic idea of the BSAS algorithm is the following : as each new point is considered, it is either assigned to an existing



(a) Real-valued global scatter plot of points $\mathbf{X}^{\mathbb{R}}(t,f)/\|\mathbf{X}^{\mathbb{R}}(t,f)\|$ weighted by their energy $\rho^2(t,f) = \|\mathbf{X}^{\mathbb{R}}(t,f)\|^2$.

(b) "Local" scatter plot of points $\hat{\mathbf{u}}(\Omega_{t,f})$ weighted by their confidence measure $\widehat{\mathcal{T}}(\Omega_{t,f})$. Neighborhood size is $|\Omega_{t,f}| = 10$.

Figure 2: Comparison of the scatter plots of points used in the standard approach and the one used by our PCA approach. STFT window size is $L = 4096$.

cluster or assigned to a newly created cluster, depending on its distance from the already formed ones. Three important factors affect the results of the BSAS algorithm :

1. the choice of the distance measure $d(\cdot,\cdot)$ between a point and a cluster;
2. the value of the threshold of dissimilarity ζ used to decide if a point is *sufficiently close* to a cluster to belong to it or not;
3. the order in which the points are presented to the algorithm.

In DEMIX the order in which the points $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$ are presented to the algorithm is determined by the confidence measure $\widehat{\mathcal{T}}(\Omega)$ of these points. The points which have the highest confidence measure are presented first. Contrary to BSAS which considers the points of the sequence one after the other, in DEMIX, when a cluster is created, all the points of the whole sequence that are considered sufficiently close to this cluster, are added to this cluster. As a consequence a point can belong to more than one cluster. The value of the threshold of dissimilarity ζ is not a fixed value in DEMIX, but an adaptative value which depends on the confidence measures of both the considered point and the point used to initially create the cluster. More formally, two directions $\mathbf{u}$, $\mathbf{u}'$ with respective confidence measure $\mathcal{T}$ and $\mathcal{T}'$ will be considered as being *sufficiently close* to each other if $d(\mathbf{u},\mathbf{u}') < \zeta(\mathcal{T},\mathcal{T}')$.

The distance measure $d(\cdot,\cdot)$ between a point and a cluster differs between DEMIX-Instantaneous and DEMIX-Anechoic, and will be detailed in due time.

As opposed to the BSAS algorithm, a further step is added at the end of the DEMIX algorithms in order to eliminate non significant clusters.

3.1 DEMIX-Instantaneous

The first step of the algorithm consists in iteratively creating clusters by selecting regions Ω_k with highest empirical confidence $\widehat{\mathcal{T}}(\Omega_k)$ and aggregating to them other regions which directions are *sufficiently close* to $\hat{\mathbf{u}}(\Omega_k)$. The number K of created clusters is determined by the algorithm and depends on the structure of the scatter plot $\{\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\}_\Omega$. The second step of the algorithm is to estimate the centroid $\hat{\mathbf{u}}_k^c$ of each cluster by first selecting a subset of *confident* points in the cluster (see section 3.1.2), then weighting these points according to their confidence value. Finally, we use a statistical test to eliminate unreliable clusters and keep $\widehat{N} \leq K$ clusters which centroids provide the estimated directions of the mixing matrix. Below we detail each step of the algorithm.

3.1.1 Cluster creation

the first step of the algorithm iteratively creates K clusters $C_k \subset P$ where P is the set of all regions Ω considered for the scatter plot.

1.1) initialize : $K = 0$, $P_K = P_0 = P$;

1.2) find the region $\Omega_K \in P_K$ with highest confidence:

$$\Omega_K := \arg \max_{\Omega \in P_K} \widehat{\mathcal{T}}(\Omega);$$

1.3) create a cluster C_K with all regions $\Omega \in P$ such that $\hat{\mathbf{u}}(\Omega)$ is *sufficiently close* to $\hat{\mathbf{u}}(\Omega_k)$;

1.4) update $P_{K+1} = P_K \setminus C_K$ by removing points of cluster C_K which are still in P_K;

1.5) stop if $P_K = \emptyset$, otherwise increment $K \leftarrow K + 1$ and go back to 1.2.

Note that in step 1.3 the newly created cluster might contain points already contained in previous clusters. Because of the sign indeterminacy in the definition of the direction, the distance between two generic unit vectors $\mathbf{u}, \mathbf{v}$, which represent source directions, needs to be carefully defined. Two directions are close to each other whenever the angle between them is small, that is to say when $|\langle \mathbf{u}, \mathbf{v} \rangle|$ is close to 1. We can therefore define the square of the direction distance by :

$$d^2(\mathbf{u}, \mathbf{v}) := \min_{|z|=1, z \in \mathbb{C}} \|\mathbf{u} - z\mathbf{v}\|^2 = 2\left(1 - |\langle \mathbf{u}, \mathbf{v} \rangle|\right). \tag{6}$$

Step 1.3 will therefore consist in including in C_K all regions such that

$$d(\hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_K)) \leq \zeta(\widehat{\mathcal{T}}(\Omega), \widehat{\mathcal{T}}(\Omega_K)) \tag{7}$$

where $\zeta(\widehat{\mathcal{T}}(\Omega), \widehat{\mathcal{T}}(\Omega_K))$ is defined in equation (8).

The adaptive threshold of dissimilarity

Now we explain how to define $\zeta(\widehat{\mathcal{T}}, \widehat{\mathcal{T}}')$. Based on the statistical model developped in Appendix A, the distance between the estimated direction $\hat{\mathbf{u}}(\Omega)$ and the "true" underlying direction of the region $\mathbf{u}(\Omega)$ satisfies :

$$\mathbb{E}\{d^2(\hat{\mathbf{u}}(\Omega), \mathbf{u}(\Omega))\} = (M-1) \cdot \sigma^2(\mathcal{T}(\Omega))$$

where $\sigma^2(\mathcal{T})$ is expressed in Eq.(27) and $\mathcal{T}(\Omega)$ is a "true" confidence value in region Ω. Since, according to the model, $\hat{\mathbf{u}}(\Omega) - \mathbf{u}(\Omega)$ and $\hat{\mathbf{u}}(\Omega_K) - \mathbf{u}(\Omega_K)$ have centered Gaussian distributions, if they are in addition assumed to be independent, we have :

$$\begin{aligned}
\mathbb{E}\{d^2(\hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_K))\} &= \mathbb{E}\{d^2(\hat{\mathbf{u}}(\Omega), \mathbf{u}(\Omega))\} \\
&\quad + d^2(\mathbf{u}(\Omega), \mathbf{u}(\Omega_K)) \\
&\quad + \mathbb{E}\{d^2(\mathbf{u}(\Omega_K), \hat{\mathbf{u}}(\Omega_K))\} \\
&= d^2(\mathbf{u}(\Omega), \mathbf{u}(\Omega_K)) \\
&\quad + (M-1)\sigma^2(\mathcal{T}(\Omega)) \\
&\quad + (M-1)\sigma^2(\mathcal{T}(\Omega_K)).
\end{aligned}$$

To ensure some robustness, we use the threshold :

$$\zeta(\mathcal{T}, \mathcal{T}') := q_2 \cdot \sqrt{M-1} \cdot \sqrt{\sigma^2(\mathcal{T}) + \sigma^2(\mathcal{T}')}, \tag{8}$$

where q_2 is a quantile which tunes the level of confidence so that the distance is compatible with the tested hypothesis $\mathbf{u}(\Omega) = \mathbf{u}(\Omega_K)$. In our experiments, we use a value of $q_2 = 2.33$ to provide a confidence level of 99 percent. In practice, every occurence of the unknown "true" confidence value $\mathcal{T}(\Omega)$ is also replaced with its empirical estimate $\widehat{\mathcal{T}}(\Omega)$ or a more robust (and more pessimistic) estimate $\tilde{\mathcal{T}}(\Omega)$ defined in Eq. (25) and depending on another quantile $q(\alpha)$.

3.1.2 Direction estimation

after creating K clusters $\{C_k\}_{k=1}^K$, we estimate their centroids $\mathbf{u}(C_k)$. Because the clusters might intersect, this estimation is based on a subset $C'_k \subset C_k$ of *confident* points that belong to C_k, and the estimation is done with the following steps.

2.1) determine the confidence threshold :

$$\eta_k := \max_{\Omega \in C_k \cap [\cup_{j \neq k} C_j]} \widehat{\mathcal{T}}(\Omega) \tag{9}$$

2.2) keep only points with sufficiently high empirical confidence value :

$$C'_k := \left\{ \Omega \in C_k \,\middle|\, \widehat{\mathcal{T}}(\Omega) \geq \eta_k \right\}.$$

2.3) estimate the centroid $\mathbf{u}(C_k)$ using equation (11) below.

Figure 3 illustrates this process.

In light of the statistical model developped in Appendix A, Eq. (26)-(28), each direction $\hat{\mathbf{u}}(\Omega)$ of the thresholded cluster C''_k is distributed as $\mathcal{N}(\mathbf{u}_k, \sigma^2(\mathcal{T}) \cdot$

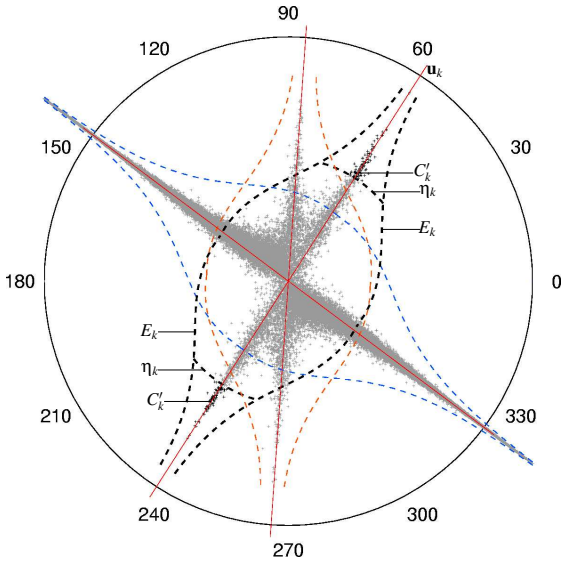


Figure 3: Illustration of how the cluster C_k is thresholded at level η_k to obtain a symmetric cluster C'_k (indicated by dark points in the scatter plot), and then estimate the direction $\mathbf{u}_k = \mathbf{u}(C_k)$. The polar scatter plot is the same as in Figure 2(b). The bold dashed line E_k indicates the enveloppe of points of cluster C_k defined by equation $d(\mathbf{u}, \hat{\mathbf{u}}(\Omega_k)) = \zeta(\mathcal{T}, \widehat{\mathcal{T}}(\Omega_k))$.

R) . The minimum variance unbiased estimator of the "true direction" $\mathbf{u}_k$ is given by :

$$\mathbf{v}_k := \frac{\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T}(\Omega)) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T}(\Omega))}$$

In practice, since the direction vectors $\hat{\mathbf{u}}(\Omega)$ are only defined up to a sign, we multiply each direction with the sign $\varepsilon(\Omega)$ such that the correlation $\langle \varepsilon(\Omega) \cdot \hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_K) \rangle$ with the direction of the point of the cluster with highest confidence is positive. Moreover the true confidence levels must be replaced with empirical estimates, which yields in practice

$$\mathbf{v}_k := \frac{\sum_{\Omega \in C'_k} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega)) \cdot \mathtt{sign}\left(\langle \hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_K) \rangle\right) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_k} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega))}. \tag{10}$$

and

$$\mathbf{u}(C_k) := \mathbf{v}_k / \|\mathbf{v}_k\| \tag{11}$$

3.1.3 Cluster Elimination

the last step of the algorithm consists in eliminating unreliable clusters. Under the model developped in Appendix A, each direction $\hat{\mathbf{u}}(\Omega)$ of a region in the cluster C'_k has the distribution $\mathcal{N}(\mathbf{u}_k, \sigma^2(\mathcal{T}(\Omega)) \cdot \mathbf{R})$ and the minimum variance

unbiased estimator defined by (10) is distributed as :

$$\mathbf{v}_k \sim \mathcal{N}\left(\mathbf{u}_k, \Big(\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T})\Big)^{-1}\mathbf{R}\right). \tag{12}$$

where $\text{trace}(\mathbf{R}) = M - 1$. To estimate the reliability of the centroid $\mathbf{u}(C_k)$ in estimating the unknown "true direction" u_k, we propose to use the expected deviation

$$\sigma^2(C_k) := \mathbb{E}\left\{\|\mathbf{v}_k - \mathbf{u}_k\|^2\right\} = (M-1) \cdot \Big(\sum_{\Omega \in C'_k} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega))\Big)^{-1} \tag{13}$$

In order to remove unreliable clusters, we now consider cluster centroids $\mathbf{u}(C_k)$ as points we want to cluster, so as to merge unreliable clusters with reliable ones. So using clusters centroids as input of our *Cluster creation* step presented in section 3.1.1, cluster C_j will merge with cluster $C_o \neq C_j$ if :

$$d\left(\mathbf{u}(C_j), \mathbf{u}(C_o)\right) < q_3 \cdot \sqrt{M-1} \cdot \sqrt{\sigma^2(C_j) + \sigma^2(C_o)}. \tag{14}$$

where q_3 is the quantile of equation (8) we use instead of q_2. In our experiments we used a value of $q_3 = 2.33 q_2$.

3.2 DEMIX-Anechoic

We now detail the DEMIX-Anechoic algorithm, which extends the DEMIX-Instantaneous algorithm to anechoic mixtures with the same main three steps. The main difference with the DEMIX-Instantaneous algorithm is in the cluster creation step, because each cluster C_k, in addition to be characterized by an *intensity profile*, is also characterised by interchannel delay(s) Δ_k. This calls for a modified definition of the centroid of a cluster. This is now a frequency dependent function $\mathbf{u}_{C_k}(f)$ parameterized by both:

- a frequency independent intensity profile $\text{abs}(\mathbf{u}_{C_k}(f))$, where the function $\text{abs}(\cdot)$ from $\mathbb{C}^M$ to $\mathbb{R}^M$ calculates the magnitude of each element of a vector.
- frequency dependent phases on each channel determined by the delays Δ_k.

The main changes in the algorithm are the new time-delay estimation step, as well as how we determine when the complex-valued direction $\hat{\mathbf{u}}(\Omega)$ of a region is *sufficiently close* to that of another region or to the centroid of a cluster.

3.2.1 Cluster creation and delay estimation

This step follow the same iterative procedure as for DEMIX-Instantaneous describe in section 3.1.1 except for step 1.3 devided now in 2 steps :

1.3.a) create a temporary cluster $\widetilde{C}_K$ with all regions $\Omega \in P_K$ with $\text{abs}(\hat{\mathbf{u}}(\Omega))$ *sufficiently close* to $\text{abs}(\hat{\mathbf{u}}(\Omega_K))$, that is regions Ω such that :

$$d(\text{abs}(\hat{\mathbf{u}}(\Omega)), \text{abs}(\hat{\mathbf{u}}(\Omega_K))) \leq q_2 \cdot \sigma(\tilde{\mathcal{T}}(\Omega_K)),$$

where $\tilde{\mathcal{T}}(\Omega_K)$ is defined in equation (25)

1.3.b) estimate the interchannel delays Δ_K for $\widetilde{C}_K$;
if Δ_K is considered as *well identified* (cf Section 4) : define the centroid $\mathbf{u}_{C_K}(f)$ using the intensity profile abs($\hat{\mathbf{u}}(\Omega_K)$) and the delays Δ_K; create the cluster C_K with all regions $\Omega \in P$ *sufficiently close* to $\mathbf{u}_{C_K}(f)$;
otherwise : reject the cluster $C_K := \tilde{C}_K$;

In Step 1.3.a, we compute distances between intensity profils. In contrast, in Step 1.3.b we need to compute the distance between a complex direction $\hat{\mathbf{u}}(\Omega)$ and a centroid direction $\mathbf{u}_{C_K}(f)$, which is frequency dependent. For example, in the stereophonic case $M = 2$, the complex direction of a given region is $\hat{\mathbf{u}} = [\cos\widehat{\theta}\ \sin\widehat{\theta} \cdot e^{i\widehat{\phi}}]^T$ while the centroidal direction is $\mathbf{u}_{C_k}(f) = [\cos\widehat{\theta}_k\ \sin\widehat{\theta}_k \cdot e^{-i2\pi\widehat{\delta}_k f}]^T$. Therefore, in Step 1.3.b we consider as *sufficiently close* all regions $\Omega \in P$ such that

$$d(\hat{\mathbf{u}}(\Omega), \mathbf{u}_{C_K}(f(\Omega))) \leq \zeta(\widehat{\mathcal{T}}(\Omega), \widehat{\mathcal{T}}(\Omega_K)), \tag{15}$$

where $f(\Omega)$ is the central frequency of the time-frequency region Ω and $\zeta(\mathcal{T}, \mathcal{T}')$ is defined in equation (8). In other words, in DEMIX-Anechoic, the distance d and the threshold ζ related to the BSAS algorithm are the same as in DEMIX-Instantaneous (see equation (7)), but in DEMIX-Anechoic, the centroid direction has to be indexed by the frequency $f(\Omega)$ of the region Ω.

3.2.2 Direction estimation

After creating the K clusters C_k, the intensity part of the centroid $\mathbf{u}_{C_k}(f)$ of those with well identified delays (cf Step 1.3.b) are updated as in the direction estimation step of the instantaneous case (see Section 3.1.2). The delays of the centroid Δ_k obtained in Step 1.3.b are preserved.

3.2.3 Cluster Elimination

The cluster elimination step is the same as in section 3.1.3, but in equation (14), instead of using the distance $d(.,.)$ between frequency independent directions, we use the following distance between frequency dependent directions :

$$d_c(\mathbf{u}_{C_i}(\cdot), \mathbf{u}_{C_j}(\cdot)) = \int d(\mathbf{u}_{C_i}(f), \mathbf{u}_{C_j}(f))df. \tag{16}$$

4 Time-Delay estimation

In this section, we present a method that estimates the time-delay of directions. As mentioned previously, if the time-delay is higher than one sample, we cannot estimate the time-delay with only one time-frequency point. It is necessary to gather several time-frequency points of the same source at different frequencies. We begin with a presentation of the approach for stereophonic mixtures, where only one delay needs to be estimated, before extending it to more channels.

4.1 Principle of the method

To explain the basic idea of the method, let us assume for a moment that only one source n is active on time frame t. Then, for each frequency the DUET ratio

satisfies $R_{21}(t,f) = \tan(\widehat{\theta}(t,f))e^{i\widehat{\phi}(t,f)} \approx \frac{a_{2n}}{a_{1n}}e^{-2i\pi f\delta_n}$, and the Inverse Fourier Transform (IFT) of $R_{21}(t,f)/|R_{21}(t,f)\| \approx e^{-2i\pi f\delta_n}$ yields a Dirac at time δ_n. Therefore, the maximum absolute value of this IFT locates the time-delay of direction n.

In practice, one rarely observe an entire time frame t where only one source is active, but as indicated in section 3.2 one can determine a set $\tilde{C}_n$ of time-frequency regions where it is likely that only one source is active. In each of the regions $\Omega \in \tilde{C}_n$, the phase of the estimated complex direction $\hat{\mathbf{u}}(\Omega)$ is $e^{i\phi(\Omega)} \approx e^{-2i\pi\delta_n f}$, where $f = f(\Omega)$ is the "central frequency" of the time-frequency region. The accuracy of this approximation is related to the value of $\sigma^2(\widehat{\mathcal{T}}(\Omega))$ as given in Equation (27). By weighting according to their precision all estimates corresponding to time-frequency regions Ω with central frequency f, one can expect to yield a more accurate estimate of the phase for a given frequency f. For that, we propose the following estimator

$$R_{\tilde{C}_n}(f) := \frac{\sum_\Omega w_f(\Omega)e^{i\widehat{\phi}(\Omega)}}{\sum_\Omega w_f(\Omega)} \approx e^{-2i\pi\delta_n f} \tag{17}$$

with

$$w_f(\Omega) := \begin{cases} 1/\sigma^2(\widehat{\mathcal{T}}(\Omega)) & \text{if } \Omega \in \widetilde{C}_n \text{ and } f = f(\Omega) \\ 0 & \text{otherwise} \end{cases} . \tag{18}$$

Eventually, the IFT of $R_{\tilde{C}_n}(f)$

$$r_{\tilde{C}_n}(\tau) := \int R_{\tilde{C}_n}(f)e^{i2\pi f\tau}df \approx \delta(\tau - \delta_n) \tag{19}$$

provides an approximation to a Dirac delta at time δ_n which precision depends on the quality of the approximation (17). The highest peak of this function provides final time-delay estimate :

$$\widehat{\delta}_n := \arg\max_\tau r_{\tilde{C}_n}(\tau) \tag{20}$$

In practice, we consider that a *well identified peak* is found if the amplitude of this main peak at $\widehat{\delta}_k$ exceeds that of all other possible peaks by at least 3dB.

Our time-delay estimator is a variant of the GCC-PHAT estimator [7], for multiple sources and a time-frequency representation instead of a single source and a frequency representation. The point of view is also different : The GCC estimator is viewed as a generalization of the cross correlation function, by the introduction of two filters for each of the the two channels, which, when properly selected, facilitates the estimation of the delay. The PHAT is one case of selected filters, developed as an ad-hoc technique to avoid the spreading of the delta function.

4.2 Delay estimation for more than two channels

For more than two channels, there are $M-1 > 1$ interchannel delays to estimate, and their definition depends on the channel we choose as reference. To avoid intrinsinc phase unstabilities when the intensity on a channel is close to zero, we propose to choose as reference the channel with largest intensity in the intensity profile $\text{abs}(\hat{\mathbf{u}}(\Omega_K)) = (u_m)_{m=1}^M$ of the highest confidence time-frequency region

used to initiate the cluster $\tilde{C}_K$: we let $m_K := \arg\max_m |u_m|$, and estimate the interchannel delays $\widehat{\delta}'_{mK}$ between each channel $m \neq m_K$ and the reference channel m_K, using the stereophonic time-delay estimation method described in section 4.1. We consider that Δ_K is *well identified* if all delays of Δ_K are *well identified*.

4.3 Discrete time implementation

In practice, time-frequency representations are only computed with a discrete grid of frequencies. As a consequence, the estimators defined in Equations (19) and (20) only provide time-delays on a discrete time grid. If the IFT of equation (19) is computed with the same frequencies as the ones used by the STFT, the resolution of this grid is one sample. It is nevertheless possible to increase this resolution by zero padding or "spectral zooming" [10] the function of equation (17).

5 Experimental study

In order to evaluate our DEMIX-Instantaneous and Anechoic methods, we propose in this section three experiments:

- to compare the proposed DEMIX-Instantaneous algorithm, with several classical clustering approaches using the ELBG algorithm and variants.
- to test the limits of the DEMIX-Instantaneous algorithm on anechoic mixtures, by varying the delay of directions from a very low value to an high value, so as to "slide" smoothly from instantaneous mixture conditions to anechoic ones.
- to compare the ability of DEMIX-Anechoic and DUET in estimating the directions of anechoic mixtures obtained by anechoic room simulations.

All experiments were performed in the stereophonic case. In addition, we also propose a comparison of the DUET[1] separation method, with the proposed frame-shifting method, by measuring separation performances on some oracle mixing matrices.

5.1 Performance measures

As our proposed DEMIX methods are able to both estimate the number of sources and the mixing directions, we propose two measures to evaluate the performance of each of these features.

5.1.1 Counting accuracy

A first measure of performance is the rate of success in the estimation of the number of sources. This measure is applied only on DEMIX, because DUET, ELBG and his variants cannot estimate the number of sources.

[1] We thank S. Rickard and C. Fearon for having graciously provided the implementation of DUET [11].

5.1.2 Accurate direction estimation

In case of success in determining the number of sources, i.e., when $\hat{N} = N$, we can also measure the *mean direction error* (MDE) which is the mean distance between true directions and estimated ones, computed with an optimized permutation to best match directions.

For a linear instantaneous mixture, given the true directions $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N]$ and estimated ones $\mathbf{A} = [\widehat{\mathbf{a}}_1 \widehat{\mathbf{a}}_2 \dots \widehat{\mathbf{a}}_N]$ the *mean direction error* (MDE) is defined as :

$$\text{MDE } (\mathbf{A}, \hat{\mathbf{A}}) := \min_{\pi \in S_N} \frac{1}{N} \sum_{n=1}^{N} d(\mathbf{a}_n, \hat{\mathbf{a}}_{\pi(n)}) \tag{21}$$

where S_N the permutation group of size N. One could also have use the *maximum direction error* :

$$\min_{\pi \in S_N} \max_{1 \leq n \leq N} d(\mathbf{a}_n, \hat{\mathbf{a}}_{\pi(n)}),$$

however our results (not presented here) with this measure were almost identical to those obtained with the MDE. To also measure the error in terms of relative precision, we also define the *relative mean direction error* (RMDE) as the MDE divided by the min-distance between true directions :

$$\text{RMDE } (\mathbf{A}, \hat{\mathbf{A}}) := \frac{\text{MDE } (\mathbf{A}, \hat{\mathbf{A}})}{\min_{n \neq n'} d(\mathbf{a}_n, \mathbf{a}_{n'})}. \tag{22}$$

The RMDE is zero if and only if the estimate is perfect, while if the RMDE is close to one, then the estimation error is of the same order of magnitude as the distance between true directions, indicating a very poor estimation quality.

We defined similar performance measures for anechoic mixtures, given $\mathbf{A}(f)$ and its estimate $\hat{\mathbf{A}}(f)$, by simply replacing $d(\cdot, \cdot)$ with the distance $d_c(., .)$ defined in Equation (16).

5.2 Evaluations on Instantaneous mixtures

We compare the DEMIX Instantaneous algorithm with ELBG, which is an improvement of the classical LBG algorithm [12], on instantaneous mixtures. We considerd four variants of the ELBG algorithm :

- ELBG on the angle data $\theta(t, f)$ obtained from the time-frequency bins $\mathbf{X}(t, f)$. That is to say the classical ELBG;
- WELBG (a *weighted* variant of ELBG) on the angle data $\theta(t, f)$ obtained from the time-frequency bins $\mathbf{X}(t, f)$ using the amplitude $\rho(t, f) = \|\mathbf{X}(t, f)\|$ as a weight;
- ELBG on the angle data $\widehat{\theta}(t, f)$ obtained from $\hat{\mathbf{u}}(t, f)$ after the PCA;
- WELBG on the PCA data $\widehat{\theta}(t, f)$, using the confidence measure $\widehat{\mathcal{T}}(t, f)$ as a weight.

The different tested algorithms are represented in the diagram of Figure 4.

For all the algorithms, we compute an STFT as a first step. We combine different scales corresponding to frame size of 64 samples to 32768 samples by a geometric step of 2. The windows used are Hanning and are applied with a

half-frame overlap. The neighborhood size used to compute the local PCA is $|\Omega| = 10$ points.

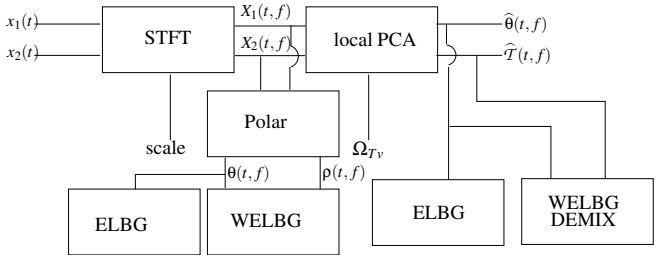


Figure 4: Block diagram of the different tested algorithms and data flow

The purpose of these four variants is to diagnose the sucess and failure of the DEMIX algorithm. In other words, we want to understand the impact on the results of: a) the "local smoothing" effect of PCA, which replaces a pointwise estimate of a direction at a given time-frequency point with a smoothed estimated averaged on a time-frequency region; b) the use of a confidence measure rather than energy to give more weight to the direction of specific time-frequency regions.

5.2.1 Experimental protocol

We apply the proposed algorithm on test signals[2] of 200 polish voice excerpts of 5 seconds sampled at 4kHz. First we study the performances of the different algorithms according to the number of sources, and second we fix the number of sources to three, and we vary the distance between these three sources.

N equally spaced sources in the first experiment, noiseless linear instantaneous mixtures were performed with mixing matrices in the most favorable shape, that is where all directions are equally spaced (as in [1]), with a number of directions going from $N = 2$ to $N = 15$.

3 sources getting closer and closer in the second experiment, 3 sources are placed with the following angles : $\theta_{l+2} = \frac{\pi}{4} + l\varepsilon\pi$, with $l \in \{-1, 0, 1\}$. In this experiment, we only vary the angular distance $\varepsilon\pi$ in order to test the robustness of the algorithm when sources get close to each other (ε small).

Since ELBG and its variants are randomly initialized, we ran them $I = 10$ times for each test mixture and focused on the smallest MDE over these 10 runs, which gives an optimistic estimate of their performance.

5.2.2 Results

For each number of sources N (respectively for each angular distance between the three sources), we choose $T = 20$ different sets of signal sources among the 200 available to build T mixtures. For each tested algorithm, we computed the

[2]These signals are those of the 2005 IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING and are available online at http://mlsp2005.conwiz.dk/

nb of sources	2	3	4	5	6
DEMIX Inst	95	100	100	95	95
nb of sources	7	8	9	10	11
DEMIX Inst	90	75	70	15	0

Figure 5: Frequency of correct count of the number of sources (in %)

counting accuracy over these T mixtures, as well as the *average* MDE among test mixtures where the number of sources was correctly counted $\widehat{N} = N$.

We observed (Figure 5) that up to $N = 8$ sources, DEMIX estimates correctly the number of directions in more than four cases out of five, but when $N > 10$ it always fails to count the number of sources. These results certainly indicate that the higher the number of sources, the less valid is the main underlying hypothesis that each source is "visible" alone in some time-frequency region.

As can be seen on Figures 6 and 7, DEMIX obtains the best performances, even if compared with the best instances of the other algorithms. As explained above, the results for DEMIX was the average RMDE on test mixtures for which the number of directions had been correctly estimated. Therefore RMDE was estimated until $N = 10$ but not beyond for DEMIX.

A remarkable fact is the behavior of DEMIX in the experiments with three sources getting very close, compared to all other algorithms. The RMDE of the four ELBG variants approaches one when the distance between true directions gets close to zero (see Figure 7), which indicates that the estimating error is nearly as high as the distance between directions: tn other words, the ELBG variants essentially confuse all directions. On the opposite, DEMIX remains very robust when the sources are very close to each other: as reported in Figure 7, the RMDE of DEMIX-Instantaneous (DEMIX Inst) remains below $3 \cdot 10^{-4}$.

We notice, by observing the results for the four variants of the ELBG, that replacing completely local direction estimates with those obtained from PCA does not significantly improve the results, while the use of the confidence measure to "boost" the most reliable directions has a much more significant impact on the performance. Yet, as can be seen on the second experiment, when a limited number of sources are present but very close, the choice of the clustering algorithm itself is important and DEMIX Instantaneous has significantly better performance.

5.3 Evaluations on synthetic anechoic mixtures

We proposed an experiment to test the limits of DEMIX-Instantaneous and the behavior of DEMIX-Anechoic as well as DUET on anechoic mixtures, by varying smoothly the degree of "anechoism" from a near instantaneous mixture to a "strong" anechoic mixture.

5.3.1 Performance measures

Since this experiment compares algorithms designed for the instantaneous and the anechoic case, it is difficult to define a simple direction distance measure which could be used to compare instantaneous directions with anechoic ones.

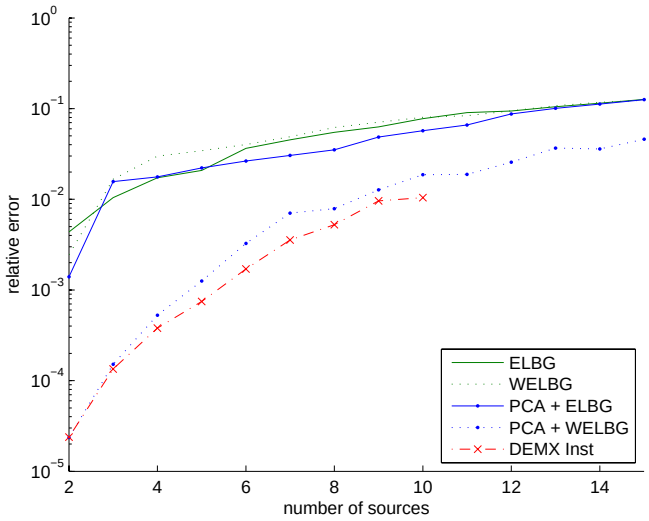


Figure 6: Relative mean direction error (RMDE) as a function of the number of sources for DEMIX-Instantaneous (DEMIX Inst) and the best instance (over 10) of the four variants of the ELBG

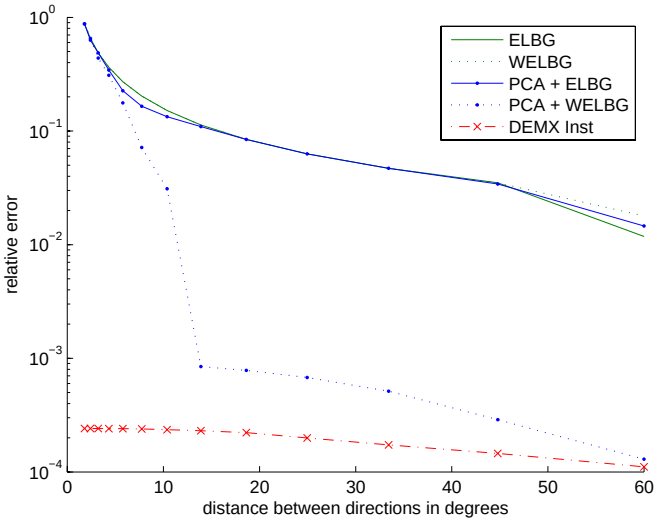


Figure 7: Relative mean direction error (RMDE) as a function of the angular distance between the 3 sources for DEMIX-Instantaneous (DEMIX Inst) and the best instance (over 10) of the four variants of the ELBG

Instead we chose to measure performance in terms of source separation quality, as measured with Sources to Distorsion Ratio (SDR) [13], using as a separation algorithm after the direction estimation step the frame-shifting (FS) method presented in appendix B. As a benchmark, we also computed these figures of merit when the separation algorithm was applied using the true mixing matrix, a method that we refered to as an oracle.

5.3.2 Experimental protocol

Similarly to the second experiment in Section 5.2.1 we used stereophonic mixtures of 3 sources with one in the middle, one on the left, and one on the right. Their ID are respectively $\theta_n \in \{\pi/2, 5\pi/12, \pi/12\}$ and the delays $\delta_n \in \{-\delta, 0, +\delta\}$. The value of the delay $\delta \geq 0$ represents the degree of "anechoism" which is varying. A frame size of 512 samples is used in the FS separation method.

5.3.3 Results

Results of the experiment are shown in Figure 8.

One can observe on Figure 8 that, while the separation performance of the classical DUET-type time-frequency masking (with oracle matrix) is comparable to that of the proposed FS method (with oracle matrix) when the delays are lower than 0.1 sample, the SDR of the FS method is systematically at least 5dB higher than that of standard time-frequency masking for delays higher than 1 sample. This is the reason why the rest of the evaluation was performed with the FS method.

The first striking observation is that for any delay δ, the performance of DEMIX-Anechoic is excellent, since it closely follows the oracle up to large delays. Indeed, even for a delay of 60 samples (150 milliseconds at 4kHz), the DEMIX Anec SDR was only 5dB below the oracle.

A second observation is that for very low delay, DEMIX-Instantaneous and DEMIX-Anechoic provide similar performance, equivalent to the oracle, and exceeding DUET by 0.5dB in terms of SDR.For a delay lower than 0.12 sample, both DEMIX-Instantaneous algorithms are no more than 0.26dB below the oracle in SDR.For delays exceeding one sample however, DEMIX-Instantaneous completely breaks down with SDR plunging more than 10dB below DEMIX-Anechoic. Indeed, for delays between 1.5 and 5 samples, DEMIX-Instantaneous performance could not even be reported for the algorithm systematically failed to count the number of sources.

As far as DUET is concerned, as already mentioned, it is not designed to estimate delays higher than one sample. With no surprise, its SDR, clearly fall when the delay exceeded 2 samples. Overall, the performance of DUET remains quite below that of DEMIX-Anechoic, and even for low delays between 10^{-2} and 10^{-1} samples, DUET SDR was 0.5dB under the ones of DEMIX-Anechoic.

5.4 Comparaison DEMIX-Anechoic / DUET on room simulated anechoic mixtures

As a third experiment, we compared the performances of the proposed DEMIX-Anechoic algorithm with the classical DUET algorithm on anechoic mixtures obtained by an anechoic room simulation with the RoomSim MATLAB toolbox [14]. The mixed sources are the same as in the previous experiments, sampled at 4kHz and of duration 5 seconds.

5.4.1 Experimental protocol

In this simulation, two cardiod microphones were placed at 20 cm from each other, and their directions crossed with a right angle. Sources were placed on

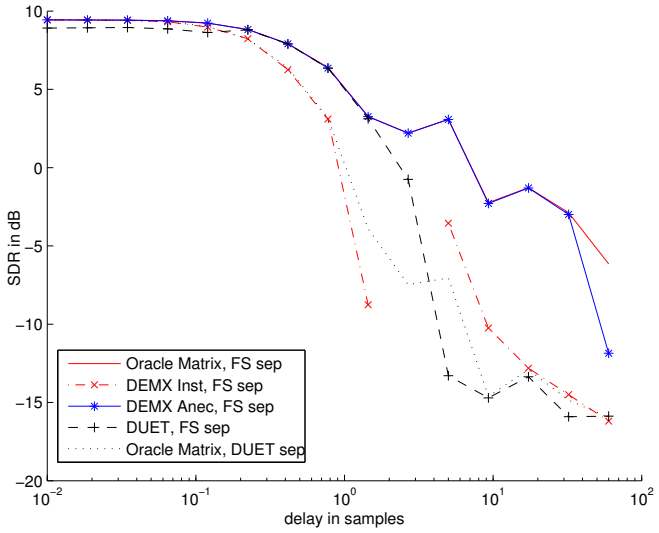


Figure 8: SDR in function of the absolute delay δ of the two side sources

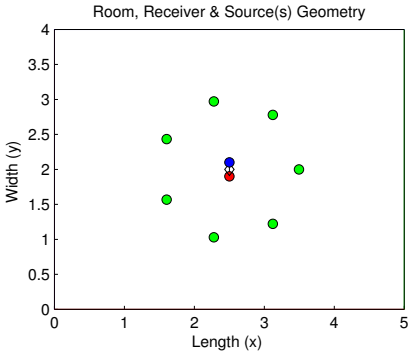


Figure 9: Room configuration for $N = 7$ sources surrounding the stereo microphone pair

n	1	2	3	4	5	6	7
θ_n	0.12	0.13	0.56	0.78	1.01	1.44	1.45
δ_n	-1	-2.2	-1.8	0	+1.8	+2.2	1

Figure 10: Intensity different θ (in radians) and delay δ (in samples) corresponding to the room configuration of Figure 9

a cercle centered in the middle of the two microphones. Sources were in the same plane as microphones, equidistant from each other, as distant as possible, and symmetric with respect to the bisector of the two microphone positions (Figure 9).

The source selection process was the same as in the first experiment of section 5.2.1. The experience consisted in estimating the performance of algorithms by changing the number of sources from $N = 2$, to $N = 7$.

nb of sources	2	3	4	5	6	7
DEMIX Anec	90	100	95	65	5	0

Figure 11: Frequency of correct count of the number of sources (in %)

5.4.2 Performance measures

A first measure of performance was the rate of success in the estimation of the number of sources. we showed that DEMIX-Anechoic estimates the number of sources until $N = 6$ (see Figure 11). Note that we cannot compare these results with DUET, because DUET doesn't estimate the number of sources and takes it as an input.

In case of success in counting the number of sources, we could also measure the average RMDE over all test mixtures.

5.4.3 Results

Figure 12 shows that the average RMDE of DEMIX-Anechoic was consistently lower than that of DUET by a factor of 10, for all considered number of sources.

DUET worked with a weighted K-Means algorithm as implemented by its authors [11]. Since the RMDE for DEMIX can only be measured when a correct number of sources is estimated, it was not computed when $N > 6$ with DEMIX-Anechoic. The RMDE for DUET was computed with the same test mixtures as those used with DEMIX-Anechoic.

nb of sources	2	3	4	5	6
DEMIX Anec	0.006	0.001	0.014	0.024	0.068
DUET	0.158	0.303	0.499	0.647	1.333

Figure 12: Average RMDE as a function of the number of sources

Two major hypotheses could explain why DEMIX-Anechoic obtains better results than DUET. First, delays exceeding one sample result in ambiguous delay estimations by phase unwrapping in DUET, an issue overcome by the GCC-PHAT type algorithm in DEMIX-Anechoic. Second, the direction of sources which are (nearly) present on only one of the channels of the mixture (i.e. with ID θ near 0 or $\pi/2$) is very unstably estimated through the ratio $R_{21}(\tau, \omega) = X_2(\tau, \omega)/X_1(\tau, \omega)$. Unlike DUET, DEMIX-Anechoic relies on a parameterization of the directions which is equally robust for all possible source directions.

6 Conclusion

We proposed a new approach to estimate the spatial directions of an unknown number of sources from a multichannel mixture in a possibly underdetermined and anechoic setting. On the experiments we conducted the proposed method was able a) to count the number of directions until 6 sources; b) to robustly estimate very close directions that classical clustering algorithms like K-Means or ELBG failed to estimate; c) to estimate delays as large as 150 samples in

simulated anechoic mixture, while performances of DUET collapse for delays higher than 2 samples.

The proposed method relies on a simple statistical model of the mixture and exploits a certain level of sparsity of the time-frequency representations to extract local estimates of the directions. Our main contribution is the use of a confidence value to robustly estimate the mixing directions, together with a method similar to GCC-PHAT to estimate the time delays of anechoic mixtures. The method seems essentially limited by the fact that it relies on the assumption that each target source significantly "emerges" from the others in sufficiently many time-frequency regions. While this is much weaker an assumption than the usual W-disjoint orthogonality underlying DUET-type methods, this condition is likely to fail when the mixture is made of many sources or the sources representations are not very sparse. One way to deal with these cases for mixtures with $M > 2$ channels would be to replace the confidence measure by a measure which indicates the likelihood that at most $M - 1$ sources are active. This would require adequate modifications of the clustering algorithm which may become significantly more complex. Another interesting perspective is to extend the present method to the convolutive case by considering sparse filters, the anechoic case being the special case of a 1-sparse filter. Yet, in the general case, the intensity difference of a direction at different frequencies would no longer be constant, and other techniques must be used to cluster directions estimated at different frequencies.

A Local statistical model

In this appendix, we analyze the relation between the (empirical) local confidence measure and the reliability of the estimated source direction based on a simple statistical model of the mixing process in a local time-frequency region.

In the instantaneous mixing model ($\delta_n = 0$), the mixing matrix $\mathbf{A}(f)$ is a constant real-valued matrix $\mathbf{A}$ which does not depend on the frequency. By taking the real or imaginary part of the complex-valued mixture model $\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f) + \mathbf{N}(t,f)$, an equivalent real-valued one is expressed as : $[\Re\mathbf{X}(t,f), \Im\mathbf{X}(t,f)] = \mathbf{A}[\Re\mathbf{S}(t,f), \Im\mathbf{S}(t,f)] + [\Re\mathbf{N}(t,f), \Im\mathbf{N}(t,f)]$.

In our local statistical model of the sources, we suppose that a source s_n, with a direction $\mathbf{a}_n$, is the most active source in the region Ω. The values of the real and imaginary parts of the STFT of this source in the region Ω are modeled as independent centered normal random variables of (large) variance σ_s^2 (which means that the STFT of the source itself is modeled as a centered circular normal complex random variable) . The contribution of the other sources, including possibly noise, are modeled by an isotropic M-dimensional centered normal distribution with covariance matrix $\sigma_n^2\mathbf{I}_M$.

Therefore, the entries $\Re\mathbf{X}(\tau,\omega), \Im\mathbf{X}(\tau,\omega)$, $(\tau,\omega) \in \Omega$ of the scatter plot $\mathbf{X}^{\mathbb{R}}(\Omega) = \mathbf{a}_n \cdot s_n^{\mathbb{R}}(\Omega) + \mathbf{N}^{\mathbb{R}}(\Omega)$ follow a centered normal distribution $\mathcal{N}(0, \Sigma_{\mathbf{X}})$ with $\Sigma_{\mathbf{X}} = \sigma_s^2\mathbf{a}_n\mathbf{a}_n^T + \sigma_n^2\mathbf{I}_M$ a real symmetric matrix. The largest eigenvalue of $\Sigma_{\mathbf{X}}$ is $\lambda_1 = \sigma_s^2 + \sigma_n^2$ associated to the principal direction $\mathbf{a}_n$, and the remaining eigenvalues are $\lambda_2 = \ldots = \lambda_M = \sigma_n^2$. It follows that the "true" direction defined by $\mathbf{a}_n$ coincides with the direction of the principal component. The "true" confidence measure defined by $\mathcal{T} := \frac{\lambda_1}{\frac{1}{M-1}\sum_{m=2}^{M}\lambda_m} = \sigma_s^2/\sigma_n^2 + 1$ can be viewed as a signal to noise ratio between the dominant source and the contribution of the other ones plus noise.

If the observation of the scatter plot $\mathbf{X}^{\mathbb{R}}(\Omega)$ were sufficient to get a perfect estimate of the covariance matrix $\Sigma_{\mathbf{X}}$, the analysis would be over. However, in practice the principal direction $\hat{\mathbf{u}}(\Omega)$ and the local confidence measure $\widehat{\mathcal{T}}(\Omega)$ are computed by PCA on sample of only $L := L(\Omega) = |\Omega|$ points, hence they only provide an estimation of the "true" direction $\mathbf{a}_n$, and an estimation of the "true" confidence $\mathcal{T}$, with a finite precision which we want to estimate, as a function of the sample size L. For that, we use results from random matrix theory. By [15, Theorem 5.7], the empirical covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{X}} := L^{-1}\mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$ follows a Wishart distribution $L^{-1}\mathcal{W}_M(\Sigma_{\mathbf{X}}, L-1)$ of dimension M. By [15, Theorem 9.4], since $\mathbf{\Sigma}_{\mathbf{X}}$ has the spectral decomposition $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{U}\Lambda\mathbf{U}^T$, as soon as the eigen-values Λ are pairwise different, the spectral decomposition of $\hat{\mathbf{\Sigma}}_{\mathbf{X}} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}^T$ converges in law, when the sample size L gets large, to a normal distribution:

$$\sqrt{L-1}\cdot(\hat{\Lambda} - \Lambda) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0, 2\Lambda^2) \tag{23}$$

$$\sqrt{L-1}\cdot(\hat{\mathbf{u}}_1 - \mathbf{u}_1) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0, \mathbf{V}) \tag{24}$$

where $\xrightarrow{\mathcal{L}}$ denotes convergence in law. The $M \times M$ covariance matrix $\mathbf{V}$ is given by

$$\begin{aligned}\mathbf{V} &= \lambda_1 \sum_{m\geq 2} \frac{\lambda_m}{(\lambda_m - \lambda_1)^2} \mathbf{u}_m \mathbf{u}_m^T \\ &= \left(\frac{\sigma_s^2}{\sigma_n^2} + 1\right) \cdot \left(\frac{\sigma_n^2}{\sigma_s^2}\right)^2 (\mathbf{I}_M - \mathbf{u}_1 \mathbf{u}_1^T) \\ &= \frac{\mathcal{T}}{(\mathcal{T}-1)^2} \cdot (\mathbf{I}_M - \mathbf{u}_1 \mathbf{u}_1^T).\end{aligned}$$

We see that the true confidence measure $\mathcal{T}$ is intrinsically related to the covariance $\mathbf{V}$ of the estimator $\hat{\mathbf{u}}(\Omega)$ of the true direction $\mathbf{a}_n$. However, in practice the "true" confidence measure is not observed but only its estimate $\hat{\mathcal{T}}$ the above relation cannot be used directly. By the asymptotic distribution of $\hat{\Lambda}$ (see Eq.(23)), denoting $\hat{\mu} := (\hat{\lambda}_1, \frac{1}{M-1}\sum_{m=2}^{M} \hat{\lambda}_m)$ and $\mu := (\sigma_s^2 + \sigma_n^2, \sigma_n^2)$ we have

$$\sqrt{L-1} \cdot (\hat{\mu} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 2 \cdot \operatorname{diag}\left(\mu_1^2, \frac{\mu_2^2}{M-1}\right)\right).$$

Writing $\frac{1}{2}\ln\hat{\mathcal{T}} = f(\hat{\mu})$ with $f(x_1, x_2) = \frac{1}{2}\ln x_1 - \frac{1}{2}\ln x_2$, by [15, Theorem 4.11] we have with $\mathbf{d} = (\left.\frac{\partial f}{\partial x_i}\right|_{\mu})_{i=1,2}$

$$\sqrt{L-1}\left(\frac{1}{2}\ln\hat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T}\right) \xrightarrow{\mathcal{L}}$$
$$\mathcal{N}\left(0, 2 \cdot \mathbf{d}^T \operatorname{diag}\left(\mu_1^2, \mu_2^2/(M-1)\right)\mathbf{d}\right).$$

One easily checks that $\mathbf{d}^T \operatorname{diag}\left(\mu_1^2, \mu_2^2/(M-1)\right)\mathbf{d} = \frac{M}{4(M-1)}$ and we obtain

$$\sqrt{\frac{2(M-1)}{M}} \cdot \sqrt{L-1}\left(\frac{1}{2}\ln\hat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

To conclude, for sufficiently large sample size L, we have

$$P\left(\mathcal{T} \leq \hat{\mathcal{T}} e^{-q\sqrt{\frac{2M}{(L-1)(M-1)}}}\right) =$$
$$P\left(\frac{1}{2}\ln\hat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T} \geq q\sqrt{\frac{M}{2(L-1)(M-1)}}\right).$$

Therefore, for any chosen confidence level $1-\alpha$ there is a quantile $q(\alpha)$ such that $\mathcal{T} \geq \widehat{\mathcal{T}} e^{-q(\alpha)\sqrt{\frac{2M}{(L-1)(M-1)}}}$ with probability exceeding $1-\alpha$. Instead of $\hat{\mathcal{T}}$, one can use the slightly more pessimistic (i.e., smaller) but more robust estimate of the local confidence measure

$$\tilde{\mathcal{T}}(\Omega) := \widehat{\mathcal{T}}(\Omega) e^{-q(\alpha)\sqrt{\frac{2M}{(L(\Omega)-1)(M-1)}}} \tag{25}$$

with $1-\alpha$ the desired confidence level. In our experiments, we choose $q(\alpha) = 2.33$, which corresponds to a confidence level of $1-\alpha = 99\%$.

We can now come back to the relation between $\tilde{\mathcal{T}}(\Omega)$ and the covariance matrix of the asymptotic distribution of $\hat{\mathbf{u}}(\Omega)$ around the true direction $\mathbf{a}_n$. By Eq. (24), the asymptotic distribution is

$$\mathcal{N}(\mathbf{a}_n, (L-1)^{-1}\mathbf{V}) = \mathcal{N}(\mathbf{a}_n, \sigma^2(\mathcal{T}) \cdot \mathbf{R}) \tag{26}$$

with

$$\sigma^2(\mathcal{T}) := \frac{\mathcal{T}}{(L-1)\cdot(\mathcal{T}-1)^2} \tag{27}$$

$$\mathbf{R} := \mathbf{I}_M - \mathbf{a}_n\mathbf{a}_n^T. \tag{28}$$

As a result, the squared distance between the estimated direction $\hat{\mathbf{u}}(\Omega)$ and the true one $\mathbf{a}_n$ is

$$\|\hat{\mathbf{u}}(\Omega) - \mathbf{a}_n\|^2 = \sigma^2(\mathcal{T}(\Omega)) \cdot \Xi$$

where the random variable $\Xi \sim \chi^2(M-1)$ is distributed according to a χ^2-distribution with $M-1$ degrees of freedom. In particular, the expected square distance is

$$\mathbb{E}\left\{\|\hat{\mathbf{u}}(\Omega) - \mathbf{a}_n\|^2\right\} = (M-1)\cdot\sigma^2(\mathcal{T}).$$

Although in the anechoic model, entries $\mathbf{X}(\tau,\omega)$, $(\tau,\omega) \in \Omega$ of the scatter plot $\mathbf{X}^{\mathbb{C}}(\Omega) = \mathbf{a}_n(f(\Omega)) \cdot s_n^{\mathbb{C}}(\Omega) + \mathbf{N}^{\mathbb{C}}(\Omega)$ follow a complex centered normal distribution, we assume that the results are the same as for the instantaneous case.

B The Frame-Shifting separation method

The Frame-Shifting (FS) method is a variant of the DUET projection method, where we shift the second channel according to the delay of the direction. Lets look at the scalar product of a direction n and a time-frequency bin $\mathbf{X}(t,f)$ defined in equation (29).

$$\begin{aligned}\langle \mathbf{a}_n(f), \mathbf{X}(t,f)\rangle = \mathbf{a}_n^H(f)\mathbf{X}(t,f) = \\ \cos(\theta_n)X_1(t,f) + \sin(\theta_n)e^{2i\pi\delta_n f}X_2(t,f)\end{aligned} \tag{29}$$

Instead of using the term $e^{2i\pi\delta_n f}$ which is equivalent of permutating circularly $x_2(\tau)$ in the frame t by a factor of δ_n, we propose to shift the window analysis $w(\tau)$ of frame t (or equivalently shift the second channel signal $x_2(\tau)$) by the δ_n factor. By this way, we avoid the unwanted side-effect of the cyclic permutation of $x_2(\tau)$. However signals are discrete, consequently we can only shift $x_2(\tau)$ by an integer sample value. So if δ_n is not an integer, we can decompose it by its round part $\lfloor\delta_n\rceil$ and its remainder part δ_n^r, so that $\delta_n = \lfloor\delta_n\rceil + \delta_n^r$. Thereby, we can shift $x_2(\tau)$ by the round part of the delay $\lfloor\delta_n\rceil$, and permutating circularly $x_2(\tau)$ by a factor of δ_n^r.

We define the following function :

$$\begin{aligned}Y\left[\theta_n, \delta_n, \mathbf{X}\right](t,f) := \\ \mathrm{STFT}\left[\cos(\theta_\mathrm{n})\mathrm{x}_1(\tau) + \sin(\theta_\mathrm{n})\mathrm{e}^{2\mathrm{i}\pi\delta_\mathrm{n}^\mathrm{r}\mathrm{f}}\mathrm{x}_2(\tau + \lfloor\delta_\mathrm{n}\rceil)\right](t,f)\end{aligned} \tag{30}$$

We first estimate the most active source n at the time-frequency point $\mathbf{X}(t,f)$ by equation (31) :

$$\hat{n}(t,f) := \arg\max_n |Y\left[\theta_n, \delta_n, \mathbf{X}\right](t,f)| \tag{31}$$

and synthesise coefficients of the sources by equation (32):

$$s_n(t,f) := \begin{cases} Y\left[\theta_n, \delta_n, \mathbf{X}\right](t,f) & if\ n = \hat{n}(t,f) \\ 0 & if\ n \neq \hat{n}(t,f) \end{cases} \tag{32}$$

References

[1] M. Z. P. Bofill, "Underdetermined blind source separation using sparse representations," in *Signal Processing*, vol. 81, no. 11, 2001, pp. 2353–2362.

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

[3] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *ICA*, 2006.

[4] ——, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," in *ICASSP 2007*, vol. 3, April 2007, pp. 745–748.

[5] Y. D. F. Abrard, "Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach," in *ISSPA 2003*. Paris, France: IEEE, July 2003.

[6] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 386–389, Mar. 2004.

[7] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics,Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[8] B. A. Paul D.O'Grady and S. T.Rickard, "Survey of sparse and non-sparse methods in source separation," *IJIST (International Journal of Imaging Systems and Technology)*, March 2005.

[9] S. Theodoridis and K. Koutroumbas, *Pattern recognition*. Academic Press, 2003.

[10] E. Hoyer and R. Stork, "The zoom fft using complex modulation," in *ICASSP*, vol. 2, May 1977, pp. 78–81.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE Transactions on Signal Processing*, vol. 52, no. 7, July 2004 2002, pp. 1830–1847.

[12] G. Patanè and M. Russo, "The enhanced LBG algorithm," *Neural Networks*, vol. 14, no. 9, pp. 1219–1237, November 2001.

[13] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 763–768.

[14] K. P. D. Campbell and G. Brown, "A matlab simulation of ?shoebox? room acoustics for use in research and teaching," *Computing and Information Systems Journal*, vol. 9, no. 3, pp. 48–51, October 2005.

[15] W. Härdel and L. Simar, Eds., *Applied multivariate statistical analysis.* Spinger-Verlag, 2003.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399