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

Simon Arberet, Rémi Gribonval, Frédéric Bimbot

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

Simon Arberet, Rémi Gribonval, *Senior Member, IEEE,* Frédéric Bimbot

***Abstract*—We propose a method to count and estimate the mixing directions in an underdetermined multichannel mixture. The approach is based on the hypothesis that in the neighbourhood of *some* time-frequency points, only one source essentially 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 statistical model to exploit a local confidence measure which detects the time-frequency regions where such 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. So as to estimate the delays of anechoic mixtures and overcome the intrinsic ambiguities of phase unwrapping as met with DUET, we propose a technique similar to GCC-PHAT which is able to estimate delays that can largely exceed one sample. We propose an extensive experimental study which shows that the resulting method is more robust in conditions where all DUET-like comparable methods fail, that is in particular : a) when time-delays largely exceed one sample; b) when the source directions are very close.**

***Index Terms*—Blind source separation, multichannel audio, delay estimation, sparse component analysis, direction of arrival**

I. Introduction

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

In this article, we consider the problem of counting and estimating the directions of the audio sources from two or more channels when there may be more sources than available channels, with an emphasis on stereophonic audio mixtures. From the mixing directions, the source signals can be estimated using the classical time-frequency masking framework [1], [2] or using more recent approaches [3]–[6]. The problem we consider is related to the *direction of arrival* (DOA) problem, which consists in estimating the physical directions of wave propagation given a controlled sensor arrangement, where the relative positions of the sensors is known. Here, we are interested in *ad hoc* sensor arrangements, thus: a) the estimated "directions" cannot necessarily be related to physical directions and should rather be considered as a parameterization of the mixture; b) it can happen that some sensors are either very close or very far apart from one another. If the sensors are very close to each other, the signals can be very similar from one sensor to another (the intensity differences are very small). On the other hand, if the sensors are far from each others, the delays between sensors can be high. In these both situations, it is difficult to estimate the source directions.

Our main contribution is a new technique to count and estimate the mixing directions which is *robust* in the sense that it can deal with the difficult settings where the underlying sensors maybe either very close to one another or very far apart. The proposed method, called DEMIX (Direction Estimation of Mixing matrIX) [7], [8] is based on a clustering algorithm which gives more weight to more reliable time-frequency regions, according to a local confidence measure similar to others proposed in the litterature [9]–[12]. An important contribution of our work is that this confidence measure is exploited via a statistical model, which is used to weight time-frequency regions in an statistically founded way, and to detect directions via a chi-square test. We also propose a new technique, which extends GGC-PHAT [13] to the multisource case and which, unlike DUET [2], is able to estimate time-delays that can largely exceed one sample.

We demonstrate with extensive experimental studies[1] the ability of our approach to count and estimate the source directions, by varying the mixing conditions (number of sources, distance between directions, level of reverberation), and by comparing our method with the classical DUET. The experiments show that the proposed DEMIX approach is able to a) blindly estimate the number of sources; up to 8 sources in the instantaneous case; and up to 5 sources in the anechoic case; b) estimate time-delays up to 100 samples; d) count and estimate nearby source directions, up to a distance of 10^{-3} degrees, with a constant relative precision better than 10^{-3}.

A. Instantaneous and Anechoic mixture model

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

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

Manuscript received July 24, 2008; revised June 29, 2009.

S. Arberet and R. Gribonval are with IRISA-INRIA, Metiss Group (Speech and Audio Processing), 35042 Rennes Cedex, France (e-mail : simon.arberet@irisa.fr, remi.gribonval@irisa.fr).

F. Bimbot is with IRISA-CNRS, Metiss Group (Speech and Audio Processing), 35042 Rennes Cedex, France (e-mail : frederic.bimbot@irisa.fr).

[1]The proposed method has been submitted to the SISEC 2008 evaluation campaign [14] for underdetermined instantaneous speech and music mixtures and obtained the best results for the mixing system estimation task.

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

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 and sign indeterminacy of the problem. If, in addition, $\delta_{mn} = 0, \forall m, n$, the mixture model is so-called *instantaneous*. Taking the Short Time Fourier Transform (STFT) $X_m(t,f)$ of each channel $x_m(\mathrm{t})$ of the mixture, the mixing process can be modeled in the time-frequency domain as $\mathbf{X}(t,f) \approx \mathbf{A}(f)\mathbf{S}(t,f) + \mathbf{N}(t,f)$ 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 a $M \times N$ complex valued 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 and are called *steering vectors (SV)* at frequency f.

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

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

The *intensity parameter (IP)* $\theta_n \in (-\pi/2, \pi/2]$ characterizes intensity difference 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 profile* defined by $\mathrm{abs}(\mathbf{a}_n(f)) \in \mathbb{R}^M$, where the function $\mathrm{abs}(\cdot)$ from $\mathbb{C}^M$ to $\mathbb{R}^M$ calculates the magnitude of each element of a vector, 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 $\delta_{1n} = 0$.

B. Related work

1) Hypothesis: 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-like methods [1], [2] rely on the assumption (called W-disjoint orthogonality) 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 [10] exploits the weaker assumption that for each source, there is *at least one* time-frequency region where this source is dominant.

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 SV $\mathbf{a}_{n(t,f)}(f)$. 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 later in this paper (see for example Figures 1 and 2). A common approach to estimate the mixing directions is thus based on a clustering algorihm applied to the points of the scatter plots.

2) Time-delay estimation: 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 IP $\theta(t,f) := \tan^{-1}|R_{21}(t,f)|$ and the delay $\delta(t,f) := -\frac{1}{2\pi f}\angle R_{21}(t,f)$ where $\angle z \in (\pi, \pi]$ is the argument of the 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 $\delta_{n(t,f)}$. Some approaches have been proposed recently to solve these issues [15]–[17] and we propose a new one in Section VI. The proposed technique is able to estimate *time delays which can largely exceed one sample*, as illustrated in Section VII-F.

3) Clustering: When it comes to actually clustering local estimates of the directions to get a global estimate of the directions, many authors 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 adjusting 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 time-frequency points with more energy, might prevent the clustering step from 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 V a new clustering algorithm which relies on the local confidence measure introduced in Section II and a statistical model described in the Section III which is used in section IV to define a proximity measure between local estimates. The proposed method is also able to count the number of sources, which is not a widely addressed task [14], [18].

II. Principle of the approach

The proposed approach to estimate the mixing directions is based on the same assumption as TIFROM [10] and relies on the use of time-frequency regions. The first step is a feature extraction step. The second step of our method is the clustering algorithm which is defined in section V.

A. 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) an estimation $\hat{\mathbf{u}}(\Omega_{t,f})$ of the SV of the most dominant source. Thus $\hat{\mathbf{u}}(\Omega_{t,f})$ is called an *estimated steering vector (ESV)*;
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 of the ESV $\hat{\mathbf{u}}(\Omega_{t,f})$, that is when essentially one source is active in this region.

As the confidence measure $\widehat{\mathcal{T}}(\Omega_{t,f})$ can discriminate the regions where essentially one source is active from the other ones, it also discriminate regions where the ESV $\hat{\mathbf{u}}(\Omega_{t,f})$ points in a direction of a SV from the regions where $\hat{\mathbf{u}}(\Omega_{t,f})$ is unlikely to point in one of the SV direction.

B. 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 $\Omega^F_{t,f}$ defined by :

$$\Omega^T_{t,f} = \{(t + kL/2, f) \quad | \quad |k| \leq K\} \tag{3}$$
$$\Omega^F_{t,f} = \{(t, f + k/L) \quad | \quad |k| \leq K\}. \tag{4}$$

where L is the STFT window size and $k \in \mathbb{Z}$.

C. 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 columns $\mathbf{X}(\tau, \omega) \in \mathbb{C}^M$, $(\tau, \omega) \in \Omega$ which will be used to analyse anechoic mixtures. For linear instantaneous mixtures, since the SV $\mathbf{a}_n(f)$ 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 columns $\Re\mathbf{X}(\tau, \omega) \in \mathbb{R}^M$ and $\Im\mathbf{X}(\tau, \omega) \in \mathbb{R}^M$, $(\tau, \omega) \in \Omega$.

D. Principal Component Analysis and confidence measure

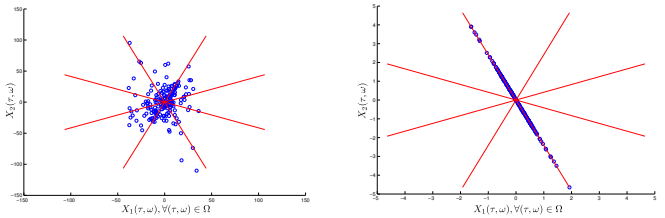
To compute the ESV $\hat{\mathbf{u}}(\Omega_{t,f})$ and their corresponding local confidence measure $\widehat{\mathcal{T}}(\Omega_{t,f})$, one can simply rely on Principal Component Analysis (PCA) applied to vectors $\mathbf{X}(\tau, \omega)$ in the region $\Omega_{t,f}$.

Performing a PCA on the local scatter plot $\mathbf{X}(\Omega)$ (resp. $\mathbf{X}^{\mathbb{R}}(\Omega)$), we obtain a *principal component (PC)* as a unit vector $\hat{\mathbf{u}}(\Omega) \in \mathbb{C}^M$ (resp. $\hat{\mathbf{u}}(\Omega) \in \mathbb{R}^M$) defined up to a multiplicative factor $e^{i\psi}$ (resp. up to a sign ± 1), 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 semi-definite matrix $\mathbf{X}(\Omega)\mathbf{X}^H(\Omega)$ (resp. the real symmetric positive semi-definite matrix $\mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$). We define the (empirical) confidence measure as :

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

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

Figure 1 shows the local scatter plot of $\mathbf{X}^{\mathbb{R}}(\Omega)$ in two time-frequency regions of an audio mixture : as expected from the theoretical results of Section III, the confidence measure is high when essentially one source is active, and low when many sources are simultaneously active.

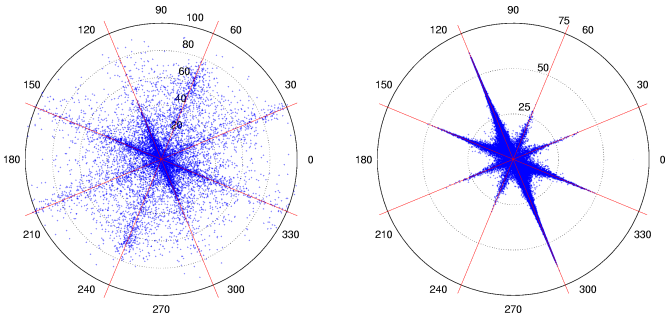


(a) Region where multiple sources contribute to the mixture. The confidence value is low (4.7 dB)

(b) Region where essentially one source contributes to the mixture. The confidence value is high (50.7 dB)

Fig. 1. Local scatter plots in two time-frequency regions. Lines indicate true source directions. STFT window size is $L = 4096$ and the size of the neighborhood is $|\Omega| = 99$.

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 mixing directions. One can see that there are many "outliers", i.e. points with high energy that are not close to the mixing directions. Conversely, Figure 2(b) displays the collection of vectors $\pm 10 \log_{10} \widehat{\mathcal{T}}(\Omega_{t,f}) \cdot \hat{\mathbf{u}}(\Omega_{t,f})$ obtained by PCA for all time-frequency regions of the signal. One can observe that points of Figure 2(b) are better concentrated along the mixing directions and thus should be better candidates to estimate the mixing direction using a clustering algorithm. This will be confirmed experimentally.



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

Fig. 2. Comparison of the scatter plots of points used in the standard approach and the ones used by the PCA approach. STFT window size is $L = 4096$.

III. Local statistical model

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

In the instantaneous mixing model ($\delta_n = 0, \forall n$), the mixing matrix $\mathbf{A}(f)$ is a 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 model 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 the proposed model, we assume that a source s_n, with a SV $\mathbf{a}_n$, 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 random variables with (large) variance σ_s^2. The contribution of the other sources, including possibly noise, are modeled by an independent, *isotropic*, M-dimensional centered 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 $\mathbf{X}^{\mathbb{R}}(\Omega) = \mathbf{a}_n \cdot S_n^{\mathbb{R}}(\Omega) + \mathbf{N}^{\mathbb{R}}(\Omega)$ have covariance :

$$\Sigma_{\mathbf{X}} = \sigma_s^2\mathbf{a}_n\mathbf{a}_n^T + \sigma_n^2\mathbf{I}_M. \tag{6}$$

The largest eigenvalue of $\Sigma_{\mathbf{X}}$ is $\lambda_1 = \sigma_s^2 + \sigma_n^2$ associated to the PC $\mathbf{u}_1$ and the remaining eigenvalues are $\lambda_2 = \ldots = \lambda_M = \sigma_n^2$. It follows that the direction of the SV $\mathbf{a}_n$ coincides with the PC $\mathbf{u}_1$. The PC $\mathbf{u}_1$ is defined up to a multiplicative factor. Nevertheless, as $\mathbf{a}_n$ is normalised, we can impose that $\mathbf{u}_1$ be normalised too and then we have $\mathbf{u}_1 = \pm\mathbf{a}_n$. Thus the ESV is defined up to a sign. The "true" confidence measure defined as :

$$\mathcal{T} := \frac{\lambda_1}{\frac{1}{M-1}\sum_{m=2}^{M}\lambda_m} = \sigma_s^2/\sigma_n^2 + 1 \tag{7}$$

can be viewed as a signal-to-noise ratio between the dominant source and the contribution of the other sources (plus noise).

A. Asymptotic distributions

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 PC $\hat{\mathbf{u}}(\Omega)$ and the local confidence measure $\widehat{\mathcal{T}}(\Omega)$ are computed on samples of only $|\Omega|$ points. Hence, $\hat{\mathbf{u}}(\Omega)$ and $\widehat{\mathcal{T}}(\Omega)$ only provide an estimation of $\mathbf{a}_n$ and $\mathcal{T}$. It is nonetheless possible to rely on random matrix theory so as to quantify the precision of these estimates as a function of the sample size $|\Omega|$.

Let $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{U}\Lambda\mathbf{U}^T$, with $\Lambda = \mathrm{diag}([\lambda_1, \cdots, \lambda_M])$, be the spectral decomposition of $\mathbf{\Sigma}_{\mathbf{X}}$, and $\hat{\mathbf{\Sigma}}_{\mathbf{X}} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}^T$, with $\hat{\Lambda} = \mathrm{diag}([\hat{\lambda}_1, \cdots, \hat{\lambda}_M])$, the one of the empirical covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{X}} := |\Omega|^{-1}\mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$.

When $\mathbf{X}^{\mathbb{R}}(\Omega)$ is Gaussian, it is known from [19, Corollary 7.2.3] that the empirical covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ follows a Wishart distribution $|\Omega|^{-1}\mathcal{W}_M(\Sigma_{\mathbf{X}}, |\Omega|-1)$ of dimension M. If $M = 2$, since the eigenvalues in Λ are pairwise distinct, it follows from [19, Theorem 13.5.1] that the following properties are satisfied :

$$\sqrt{|\Omega|-1}\cdot(\hat{\lambda}_1 - \lambda_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\lambda_1^2) \tag{H_1}$$

$$\sqrt{|\Omega|-1}\cdot\sum_{m=2}^{M}(\hat{\lambda}_m - \lambda_m) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 2\sum_{m=2}^{M}\lambda_m^2\right) \tag{H_2}$$

$$\sqrt{|\Omega|-1}\cdot(\hat{\mathbf{u}}_1 - \mathbf{u}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{V}) \tag{H_3}$$

where $\xrightarrow{\mathcal{L}}$ denotes convergence in law when the sample size $|\Omega|$ gets large. The $M \times M$ covariance matrix $\mathbf{V}$ is given by :

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

The same properties can be proved [20] when $M > 2$ and $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_M$. From now, we will assume that H_1, H_2, H_3 hold, which is likely to hold for a much wider class of distributions of $\mathbf{X}^{\mathbb{R}}(\Omega)$ than the only Gaussian distribution. Discussing the full extent of the validity of H_1, H_2, H_3 is beyond the scope of this paper.

Although the convergences in H_1, H_2, H_3 are asymptotic (for a large sample size $|\Omega|$), we will use a pretty small sample size ($|\Omega| = 10$). Experimental results in section VII show that the method is nonetheless working well.

B. Robust confidence measure

So as to evaluate the quality of the ESV using the asymptotic relation of hypothesis H_3, we define a measure called *robust empirical confidence*, which is a bound that has a high probability to lie below the true (unknown) confidence measure.

Definition *(Robust empirical confidence)* Using the empirical confidence measure $\widehat{\mathcal{T}}(\Omega)$, we define the robust empirical confidence $\tilde{\mathcal{T}}(\Omega)$ of level $1-\alpha$ by :

$$\tilde{\mathcal{T}}(\Omega) := \widehat{\mathcal{T}}(\Omega)e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}} \tag{8}$$

where $q(\alpha)$ is the quantile of order $1-\alpha$ of the Gaussian law with zero mean and variance one.

The quantile $q(\alpha) = F^{-1}(1-\alpha)$ is defined by the inverse function F^{-1} of the cumulative distribution function (CDF) $F(q) = \int_{-\infty}^{q}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$ of law $\mathcal{N}(0,1)$.

Thanks to Theorem 3.1 below, for a fixed and sufficiently large sample size $|\Omega|$, the value of the "true" confidence measure $\mathcal{T}(\Omega)$ is guaranteed to be higher than the one of the robust empirical confidence $\tilde{\mathcal{T}}(\Omega)$ with a probability $1-\alpha$. In our experiments, we chose $q(\alpha) = 6.3$.

Theorem 3.1: (Confidence interval for the confidence measure) Assume that H_1, H_2 hold, and let $\mathcal{T}$ be defined by Equation (7) and $\widehat{\mathcal{T}} := \frac{\hat{\lambda}_1}{\frac{1}{M-1}\sum_{m=2}^{M}\hat{\lambda}_m}$. Then :

$$P\left(\mathcal{T} \geq \widehat{\mathcal{T}} e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}}\right) \xrightarrow[|\Omega|\to\infty]{} 1-\alpha \quad (9)$$

where $q(\alpha)$ is the quantile of level $1-\alpha$ of law $\mathcal{N}(0,1)$.

Proof: From H_1 and H_2, 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 :

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

Writing $\frac{1}{2}\ln\hat{\mathcal{T}} = f(\hat{\mu})$ with $f(x_1,x_2) = \frac{1}{2}\ln x_1 - \frac{1}{2}\ln x_2$, from [19, Theorem 4.2.3] we have with $\mathbf{d} = (\frac{\partial f}{\partial x_i}\Big|_{\mu})_{i=1,2}$:

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

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

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

To conclude, for sufficiently large sample size $|\Omega|$, we have :

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

where $F(q)$ is the CDF of law $\mathcal{N}(0,1)$. Then, there is a quantile $q(\alpha) = F^{-1}(1-\alpha)$ such that $\mathcal{T} \geq \widehat{\mathcal{T}} e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}}$ with probability exceeding $1-\alpha$. ■

C. Precision of the direction estimation

Let us now come back to the relation between the above defined $\tilde{\mathcal{T}}(\Omega)$ and the covariance matrix $\mathbf{V}$ of the asymptotic distribution of $\hat{\mathbf{u}}(\Omega)$ around the SV $\mathbf{a}_n$.

Proposition 3.2: If the mixture covariance matrix $\Sigma_{\mathbf{X}}$ is defined as in Equation (6) and if hypothesis (H_3) holds, then the distribution of the PC $\hat{\mathbf{u}}_1$ converges in law, when the sample size $|\Omega|$ is large, to :

$$\hat{\mathbf{u}}_1 \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{a}_n, \sigma^2(\mathcal{T}(\Omega))\cdot\mathbf{R}) \quad (10)$$

with :

$$\sigma^2(\mathcal{T}) := \frac{\mathcal{T}}{(|\Omega|-1)\cdot(\mathcal{T}-1)^2} \quad (11)$$
$$\mathbf{R} := \mathbf{I}_M - \mathbf{a}_n\mathbf{a}_n^T. \quad (12)$$

and also :

$$\mathbb{E}\left\{\|\hat{\mathbf{u}}_1 - \mathbf{a}_n\|^2\right\} = (M-1)\cdot\sigma^2(\mathcal{T}). \quad (13)$$

Proof: According to H_3, the asymptotic distribution of $\hat{\mathbf{u}}_1$ is :

$$\hat{\mathbf{u}}_1 \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{a}_n, (|\Omega|-1)^{-1}\mathbf{V})$$

Also, from Equation (10), the square of the asymptotic distance between $\hat{\mathbf{u}}_1$ and $\mathbf{a}_n$ is :

$$\|\hat{\mathbf{u}}_1 - \mathbf{a}_n\|^2 = \sigma^2(\mathcal{T}(\Omega))\cdot\Xi$$

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

D. χ^2 test to decide if two regions correspond to the same direction

We assume two independent random variables $\hat{\mathbf{u}}(\Omega_1)$, $\hat{\mathbf{u}}(\Omega_2)$ distributed according to the following laws :

$$\hat{\mathbf{u}}(\Omega_1) \sim \mathcal{N}(\mathbf{a}_n, \sigma^2(\mathcal{T}(\Omega_1))\cdot\mathbf{R}) \quad (14)$$
$$\hat{\mathbf{u}}(\Omega_2) \sim \mathcal{N}(\mathbf{a}_{n'}, \sigma^2(\mathcal{T}(\Omega_2))\cdot\mathbf{R}) \quad (15)$$

We want to test the hypothesis $\mathcal{H}_0 = \{\mathbf{a}_n = \mathbf{a}_{n'}\}$ against the hypothesis $\mathcal{H}_1 = \{\mathbf{a}_n \neq \mathbf{a}_{n'}\}$. If $\mathbf{a}_n = \mathbf{a}_{n'}$, the square distance between $\hat{\mathbf{u}}(\Omega_1)$ and $\hat{\mathbf{u}}(\Omega_2)$ is given by :

$$\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2 = \left(\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))\right)\Xi \quad (16)$$

where $\Xi \sim \chi^2(M-1)$. Thus :

$$P\left(\frac{\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2}{\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))} \geq q_{\chi^2_{M-1}}(\alpha)\right) = \alpha$$

where $q_{\chi^2_{M-1}}(\alpha)$ is the quantile of level α of the χ^2 law with $M-1$ degrees of freedom. Thus, the χ^2 test of level α consists in rejecting hypothesis $\mathcal{H}_0$ if :

$$\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2 \geq q_{\chi^2_{M-1}}(\alpha)\cdot\left(\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))\right) \quad (17)$$

IV. Proximity measure

Before using a clustering algorithm for estimating the mixture directions, we need to define a *proximity measure* so as to quantify the degree of similarity between pairs $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$. As these pairs are indexed by a time-frequency region Ω, we use the term *region* to indicate these pairs. We first define the distance measure between two SV $\mathbf{u}_1$ and $\mathbf{u}_2$, and then we use the asymptotic results from section III to define the distance measure between two regions $\left(\hat{\mathbf{u}}(\Omega_1), \widehat{\mathcal{T}}(\Omega_1)\right)$ and $\left(\hat{\mathbf{u}}(\Omega_2), \widehat{\mathcal{T}}(\Omega_2)\right)$.

A. Distance measure between two steering vector (SV)

Because of the sign indeterminacy (or more generally, in the complex case, the phase indeterminacy) of the ESV, we define the square of the direction distance measure between two unit SV $\mathbf{u}_1$ and $\mathbf{u}_2$ by :

$$d^2(\mathbf{u}_1, \mathbf{u}_2) := \min_{|z|=1, z\in\mathbb{C}} \|\mathbf{u}_1 - z\mathbf{u}_2\|^2 = 2\left(1 - |\langle\mathbf{u}_1, \mathbf{u}_2\rangle|\right). \quad (18)$$

As an exemple, let $\mathbf{u}_1$ and $\mathbf{u}_2$ be two real 2-dimensional SV, with IP θ_1 and θ_2 and such that $\theta_1 - \theta_2 \approx 0$. Then

$$d^2(\mathbf{u}_1, \mathbf{u}_2) = 2\left(1 - |\cos(\theta_1 - \theta_2)|\right) \approx (\theta_1 - \theta_2)^2.$$

B. Distance measure between two regions

According to the model developped in section III, two regions Ω_1 and Ω_2 belong to the same cluster if the asymptotic Gaussian distribution of their respective ESV $\hat{\mathbf{u}}(\Omega_1)$ and $\hat{\mathbf{u}}(\Omega_2)$ have the same mean, that is $\mathbf{u}(\Omega_1) = \mathbf{u}(\Omega_2)$. The χ^2 test of Equation (17) can be used to test this hypothesis.

Define $\mathrm{D}_{\mathcal{T}}$ as :

$$\mathrm{D}_{\mathcal{T}}\left((\mathbf{u}_1, \mathcal{T}_1), (\mathbf{u}_2, \mathcal{T}_2)\right) := \frac{d(\mathbf{u}_1, \mathbf{u}_2)}{\sqrt{\sigma^2(\mathcal{T}_1) + \sigma^2(\mathcal{T}_2)}}, \qquad (19)$$

and let $\zeta = \sqrt{q_{\chi^2_{M-1}}(\alpha)}$. The hypothesis $\mathbf{u}(\Omega_1) = \mathbf{u}(\Omega_2)$ is accepted (with confidence level α) if :

$$\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(\Omega_1), \widehat{\mathcal{T}}(\Omega_1)\right), \left(\hat{\mathbf{u}}(\Omega_2), \widehat{\mathcal{T}}(\Omega_2)\right)\right) \leq \zeta \qquad (20)$$

In other words, if (20) holds then $\hat{\mathbf{u}}(\Omega_1)$ and $\hat{\mathbf{u}}(\Omega_2)$ are considered as *sufficiently close* to one another to be merged.

This distance measure $\mathrm{D}_{\mathcal{T}}$ is similar to the known Fisher criterion function [21]. In practice the occurrence of the unknown "true" confidence values $\mathcal{T}(\Omega)$ are not accessible. Thus, we replace them by their empirical estimate $\widehat{\mathcal{T}}(\Omega)$, or a more pessimistic estimate $\tilde{\mathcal{T}}(\Omega) < \widehat{\mathcal{T}}(\Omega)$ defined in Equation (8) and depending on quantile $q(\alpha)$.

We will see in the next section how the distance measure $\mathrm{D}_{\mathcal{T}}$, obtained by our statistical model, is used to cluster the directions, and also (on Figure 3) how the shape of the clusters, determined by measure $\mathrm{D}_{\mathcal{T}}$, is matching with the data.

V. Direction estimation with a clustering Algorithm

In this section we describe the two proposed DEMIX (Direction Estimation of Mixing matrIX) algorithms used to estimate the mixture directions by combining and clustering the local estimates $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$.

According to the classification of Theodoridis and Koutroumbas [22], there are three families of clustering algorithms : 1) partitional clustering based on a cost function, 2) hierarchical clustering, 3) sequential clustering. We begin by a brief review of the principles and pros and cons of these families, before describing our algorithms which are related to the BSAS algorithm [22], of the third family.

A. Classification of clustering algorithms

1) Partitional clustering based on a cost function: We could do the clustering with a standard algorithm like the K-means (also known as LBG [23]), which minimises iteratively the within-class variance. But this approach has two main drawbacks :

- There is no guarantee to converge to a global optimum, and the obtained centroids can depend on the initial partitions. Thus the general strategy for the problem, as it is actually done by DUET, is to run the algorithm many times with random initial partitions. Another solution adopted by the ELBG [24] is to call for a "roulette" mechanism typical of genetic algorithms so as to avoid convergence to a local optimum. We will evaluate the performance of this last approach in section VII.
- *The number N of clusters must be provided as an input parameter.* A large number of attempts have been made to estimate the appropriate N, but their performance are often data dependent [25] and/or computationally intensive [18].

2) Hierarchical clustering: Hierarchical clustering (HC) algorithms organise data into a hierarchical structure according to the proximity matrix [25]. There are some heuristics to estimate the number N of clusters, but the main problem of HC algorithms is that the *computational complexity* is generally at least $O(Q^2)$, which is a limit for our application where the number Q of data points is generally larger than one million.

3) Sequential Clustering: These algorithms produce a single clustering [22] and are thus faster. Algorithm like the Basic Sequential Algorithmic Scheme (BSAS) [22], does not know a priori the number of clusters and thus can be used to count the number N of sources.

B. DEMIX-Instantaneous

First we present DEMIX-Instantaneous, which is designed for instantaneous mixtures, then we present in section V-C DEMIX-Anechoic for anechoic mixtures.

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 ESV are *sufficiently close* to $\hat{\mathbf{u}}(\Omega_k)$, in the sense of Eq. (20)). The number K of clusters thus created depends on the structure of the scatter plot of all the regions $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$ (see Figure 2). 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* regions and then weighting these regions according to their confidence value. As some clusters may be created by some outliers, we finally 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.

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 of the scatter plot. As each cluster corresponds to an estimated mixture direction, the aim of this step is also to get a first estimation of the number of sources.

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 regions 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 may contain regions already contained in previous clusters.

2) Direction estimation : after creating K clusters $\{C_k\}_{k=1}^K$, we estimate their centroids $\mathbf{u}(C_k)$. As we can see on Figure 3, the distribution of the points/regions around a mixture direction k is "symmetrical" only when the confidence measure is large enough. Thus, so as to have non biased estimation for direction k, the estimation is based on a subset $C'_k \subset C_k$ of *confident* regions which have a confidence measure larger than an adaptative threshold. To define this threshold, we first set the condition that a region cannot be used to estimate two different directions; secondly, so as to have a maximum number of regions to estimate a direction, we force the threshold to be as small as possible.

The estimation is finally done in the following steps :

2.1) determine the confidence threshold :

$$\eta_k := \max_{\Omega \in C_k \cap [\cup_{j \neq k} C_j]} \widehat{\mathcal{T}}(\Omega) \tag{21}$$

2.2) keep only regions with sufficiently high empirical confidence values (the other ones are no longer used in the rest of the algorithm) :

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

Figure 3 illustrates this process.

2.3) estimate the centroid $\mathbf{u}(C_k)$ using the regions from cluster C'_k.

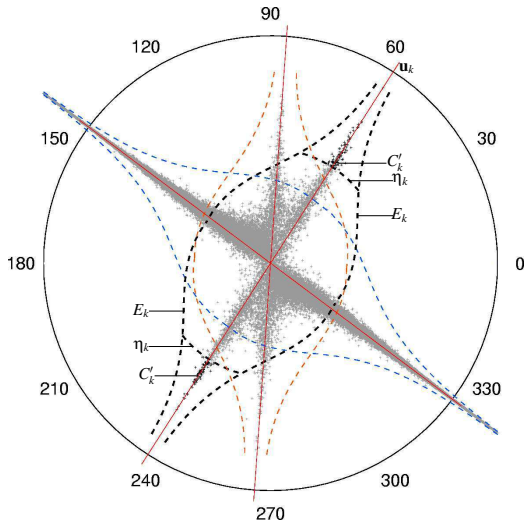


Fig. 3. Illustration of how the cluster C_k is thresholded at level η_k to obtain a symmetric cluster C'_k (indicated by dark points in the scatter plot), and then estimate the direction $\mathbf{u}_k$. The polar scatter plot is the same as in Figure 2(b) but for a different mixture. The bold dashed line E_k indicates the envelope of cluster C_k defined by points $(\mathbf{u}, \mathcal{T})$ such that $\mathrm{D}_{\mathcal{T}}\left((\mathbf{u}, \mathcal{T}), \left(\hat{\mathbf{u}}(\Omega_k), \widehat{\mathcal{T}}(\Omega_k)\right)\right) = \zeta$

In light of the statistical model developed in Section III, Eq. (10)-(12), each ESV $\hat{\mathbf{u}}(\Omega)$ of the thresholded cluster C'_k is distributed as $\mathcal{N}(\mathbf{u}_k, \sigma^2(\mathcal{T}) \cdot \mathbf{R})$. The minimum variance unbiased estimator of the "true direction" $\mathbf{u}_k$ is given by :

$$\mathbf{v}_k := \frac{\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T}(\Omega)) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T}(\Omega))} \tag{22}$$

In practice, since the ESV $\hat{\mathbf{u}}(\Omega)$ 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 $\hat{\mathbf{u}}(\Omega_K)$ is positive. Moreover, the true confidence levels must be replaced with empirical estimates :

$$\tilde{\mathbf{v}}_k := \frac{\sum_{\Omega \in C'_k} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega)) \cdot \mathtt{sign}\left(\langle \hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_K) \rangle\right) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_k} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega))}. \tag{23}$$

Finally, the direction estimation formula for cluster C_k after normalisation is :

$$\mathbf{u}(C_k) := \tilde{\mathbf{v}}_k / \|\tilde{\mathbf{v}}_k\| \tag{24}$$

3) Cluster Elimination : the last step of the algorithm consists in eliminating unreliable clusters. When two clusters are very close, it may be the consequence of two nearby directions, or one of these clusters may have been created because of some outliers. However, if we can define a confidence measure for a cluster, we can assume that two clusters which are close to each other are associated with two distinct directions, if and only if these two clusters have a large confidence measure.

Under the model developed in section III, the minimum variance unbiased estimator defined by (22) is distributed as :

$$\mathbf{v}_k \sim \mathcal{N}\left(\mathbf{u}_k, \Big(\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T})\Big)^{-1} \mathbf{R}\right). \tag{25}$$

Thanks to Proposition 3.2, the error of the estimation $\mathbf{v}_k$ of direction $\mathbf{u}_k$ is characterized by :

$$\sigma^2(C_k) := \mathbb{E}\left\{\|\mathbf{v}_k - \mathbf{u}_k\|^2\right\} = (M-1) \cdot \Big(\sum_{\Omega \in C'_k} \sigma^{-2}(\mathcal{T}(\Omega))\Big)^{-1} \tag{26}$$

In practice, since the "true" confidence measure $\mathcal{T}(\Omega)$ is not known, we replace it by the value of the robust empirical confidence measure $\tilde{\mathcal{T}}(\Omega)$. The variance measure $\sigma^2(C_k)$ can be converted into a confidence measure using the inverse function of Equation (11).

The cluster elimination step consists in reiterating the *Cluster creation* step of section V-B1, with cluster centroids $\mathbf{u}(C_k)$ and their associated confidence measure $\mathcal{T}(C_k)$ as input instead of regions $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$.

To summarize, cluster C_j will merge with cluster $C_o \neq C_j$ if :

$$\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(C_j), \widehat{\mathcal{T}}(C_j)\right), \left(\hat{\mathbf{u}}(C_o), \widehat{\mathcal{T}}(C_o)\right)\right) \leq \zeta_c \tag{27}$$

The value of ζ_c has to be larger than the threshlold ζ of Eq. (20) to efficiently eliminate unreliable clusters. Notice that if we knew (a priori) the number N of directions, we could implement this *cluster elimination* step by keeping the N directions which have the largest confidence value $\widehat{\mathcal{T}}(C_j)$.

C. DEMIX-Anechoic

We now detail the DEMIX-Anechoic algorithm. The main difference with the DEMIX-Instantaneous algorithm lies in the cluster creation step, as each mixture direction, in addition to be characterized by an *intensity profile* $\mathrm{abs}\,(\mathbf{a}_n(f)) \in \mathbb{R}^M$, is also characterised by interchannel delays Δ_n. Thus the

centroid of a cluster is now defined by a frequency dependent steering vector $\mathbf{u}_{C_k}(f)$ parameterized by both :

- a frequency independent intensity profile $\mathrm{abs}(\mathbf{u}_{C_k}(f))$.
- frequency dependent phases on each channel determined by the delays $\widehat{\Delta}_k$.

The main changes in the algorithm are: a) the incorporation of the time-delay estimation step, b) the test to determine when the complex-valued ESV $\hat{\mathbf{u}}(\Omega)$ of a region is *sufficiently close* to a cluster.

1) Cluster creation and delay estimation : This step follows the same iterative procedure as for DEMIX-Instantaneous as described in section V-B1, except for step 1.3 divided now in 2 steps :

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

$$d(\mathrm{abs}(\hat{\mathbf{u}}(\Omega)), \mathrm{abs}(\hat{\mathbf{u}}(\Omega_K))) \leq \zeta_2 \cdot \sigma(\tilde{\mathcal{T}}(\Omega_K)),$$

where $\tilde{\mathcal{T}}(\Omega_K)$ is defined in Equation (8) and ζ_2 is a threshold.

1.3.b) estimate the interchannel delays $\widehat{\Delta}_K$ for $\widetilde{C}_K$;
if $\widehat{\Delta}_K$ is considered as *well identified* (cf Section VI) : define the centroid $\mathbf{u}_{C_K}(f)$ using the intensity profile $\mathrm{abs}(\hat{\mathbf{u}}(\Omega_K))$ and the delays $\widehat{\Delta}_K$; 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.b, we need to compute the distance between an ESV $\hat{\mathbf{u}}(\Omega)$ and the centroid SV $\mathbf{u}_{C_K}(f)$, which is frequency dependent. Therefore, we consider as *sufficiently close* all regions $\Omega \in P$ such that :

$$\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right), \left(\mathbf{u}_{C_K}(f(\Omega)), \widehat{\mathcal{T}}(\Omega_K)\right)\right) \leq \zeta, \quad (28)$$

where $f(\Omega)$ is the central frequency of the time-frequency region Ω.

2) Direction estimation and cluster elimination: The *direction estimation* step and the *cluster elimination* step are similar to the last two steps of DEMIX-Instantaneous. The main difference is that, instead of using the distance $d(.,.)$ in the definition of $\mathrm{D}_{\mathcal{T}}$ (see Equation (19)), we use the distance :

$$d_c(\mathbf{u}_{C_i}(\cdot), \mathbf{u}_{C_j}(\cdot)) = \int d(\mathbf{u}_{C_i}(f), \mathbf{u}_{C_j}(f)) df. \quad (29)$$

so as to take into account the fact that the steering vectors depend on the frequency f.

VI. Time-Delay estimation

In this section, we present a method that estimates the time-delay of directions. 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.

A. Principle of the method

1) Case where only one source is active: To explain the basic idea of the method, let us assume for a moment that only one source n is active in 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 :

$$r_{21}(\tau) := \int \frac{R_{21}(t,f)}{|R_{21}(t,f)|} e^{i2\pi f \tau} df \approx \delta(\tau - \delta_n) \quad (30)$$

The GCC-PHAT method [13] consists in detecting the peak in function $r_{21}(\tau)$ with the following estimator :

$$\widehat{\delta}_n := \arg\max_{\tau} r_{21}(\tau) \quad (31)$$

2) Case where more than one source are active: In practice, one rarely observes an entire time frame t where only one source is active, but as indicated in section V-C1 one can determine a set $\tilde{C}_n$ of time-frequency regions, which have similar intensity profiles, and where it is likely that only one source is active. In each of the regions $\Omega \in \tilde{C}_n$, the phase of the ESV $\hat{\mathbf{u}}(\Omega)$ 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 (11). One can expect to obtain a more accurate estimate of the phase for a given frequency f, by weighting all estimates corresponding to time-frequency regions Ω with central frequency f according to their precision. For that purpose, we propose the following estimator :

$$R_{\tilde{C}_n}(f) := \frac{\sum_\Omega w_f(\Omega) e^{i\widehat{\phi}(\Omega)}}{\sum_\Omega w_f(\Omega)} \approx e^{-2i\pi\delta_n f} \quad (32)$$

with

$$w_f(\Omega) := \begin{cases} 1/\sigma^2(\widehat{\mathcal{T}}(\Omega)) & \text{if } \Omega \in \widetilde{C}_n \text{ and } f = f(\Omega) \\ 0 & \text{otherwise} \end{cases}. \quad (33)$$

The IFT of $R_{\tilde{C}_n}(f)$ is expected to be approximately a Dirac delta at time δ_n.

$$r_{\tilde{C}_n}(\tau) := \int R_{\tilde{C}_n}(f) e^{i2\pi f \tau} df \approx \delta(\tau - \delta_n) \quad (34)$$

The highest peak of this function provides the final time-delay estimate :

$$\widehat{\delta}_n := \arg\max_{\tau} r_{\tilde{C}_n}(\tau) \quad (35)$$

In practice, we consider that a *well identified* peak is found if the amplitude $r_{\tilde{C}_n}(\widehat{\delta}_n)$ of the main peak of function $r_{\tilde{C}_n}(\tau)$ exceeds that of all other possible peaks by at least 3dB. Notice that our time-delay estimator extends the GCC-PHAT estimator [13] to the case of multiple sources.

B. 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 a reference. Notice that, even for a time-frequency point where only one source is active, if the intensity on a channel m is close to zero (that is $\Re\mathrm{e}X_m(t,f) \approx 0$ and $\Im\mathrm{m}X_m(t,f) \approx 0$), then the phase estimation $\angle X_m(t,f) = \tan^{-1}\left(\frac{\Im\mathrm{m}X_m(t,f)}{\Re\mathrm{e}X_m(t,f)}\right)$ on that channel is unstable, as well as the phase difference between channel m and a channel $k \neq m$. To avoid these intrinsic phase

unstabilities, we propose to choose as a reference, for cluster C_K, the channel with the largest intensity in the intensity profile $\mathrm{abs}(\hat{\mathbf{u}}(\Omega_K)) = (u_m)_{m=1}^M$ of region Ω_K: we let $m_K := \arg\max_m |u_m|$ and estimate the interchannel delays $\widehat{\delta}_{m,mK}$ between each channel $m \neq m_K$ and the reference channel m_K, using the stereophonic time-delay estimation method described in section VI-A. We consider that $\widehat{\Delta}_K$ is *well identified* if all delays of $\widehat{\Delta}_K$ are *well identified*.

C. 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 (34) and (35) only provide time-delays on a discrete time grid. If the IFT of Equation (34) is computed with the same frequencies as those used by the STFT $\mathbf{X}(t,f)$, then the temporal resolution of the delay estimator of Equation (35) is one sample. It is nevertheless possible to increase this resolution by zero padding or "spectral zooming" [26] the function of Equation (32).

VII. Experimental study

In order to evaluate the DEMIX-Instantaneous and Anechoic methods, we propose in this section experiments aiming at :

- comparing the proposed DEMIX-Instantaneous algorithm with classical clustering approaches using the (K-Means like) ELBG algorithm and variants.
- testing the limits of the DEMIX-Instantaneous algorithm on anechoic mixtures.
- comparing the ability of DEMIX-Anechoic and DUET[2] in estimating the directions of anechoic and convolutive mixtures.

All experiments were performed in the stereophonic case. Thus the SV are defined according to Equation (2), and in the instantaneous case the ESV are only parametrized by an IP $\widehat{\theta}(t,f)$.

A. Clustering algorithm variants

The purpose of these four variants of ELBG is to diagnose the success 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 an IP at a given time-frequency point with a smoothed estimate averaged on a time-frequency region; b) the use of a confidence measure rather than the energy value to give more weight to the IP of specific time-frequency region.

We compare the DEMIX-Instantaneous algorithm with ELBG [24], which is an improvement of the classical LBG (i.e. K-means) algorithm [23], on instantaneous mixtures. We considered four variants of the ELBG algorithm :

- ELBG on the IP $\theta(t,f)$ 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 IP $\theta(t,f)$ 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 IP $\widehat{\theta}(t,f)$ obtained from the ESV $\hat{\mathbf{u}}(t,f)$ after the PCA;
- WELBG on the IP $\widehat{\theta}(t,f)$, using the optimal weight $1 \Big/ \sigma^2\left(\widehat{\mathcal{T}}(t,f)\right)$ where $\sigma^2(\mathcal{T})$ is defined in Equation (11).

The different algorithms tested in this study are represented on Figure 4.

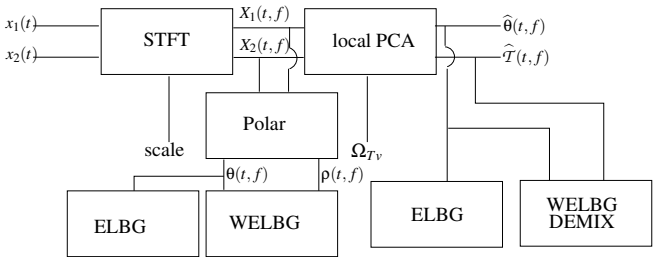


Fig. 4. Block diagram of the different tested algorithms and data flow

We also propose one variant of DEMIX-Instantaneous where, instead of estimating the mixing direction with a weighted mean using Equations (23) and (24), we take the "best region", that is the centroid $\hat{\mathbf{u}}(\Omega_K)$ as a direction estimation (we call this variant DEMIX-Instantaneous-BR).

B. Performance measures

As the 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 them.

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 its variants do not estimate the number of sources.

2) Accurate directions estimation: We propose a performance measure called 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 $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1 \widehat{\mathbf{a}}_2 \dots \widehat{\mathbf{a}}_N]$ the *mean direction error* (MDE) is defined as :

$$\text{MDE } (\mathbf{A}, \hat{\mathbf{A}}) := \min_{\pi \in S_N} \frac{1}{N} \sum_{n=1}^{N} d(\mathbf{a}_n, \hat{\mathbf{a}}_{\pi(n)}) \qquad (36)$$

where S_N is the permutation group of size N.

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 :

$$\text{RMDE } (\mathbf{A}, \hat{\mathbf{A}}) := \frac{\text{MDE } (\mathbf{A}, \hat{\mathbf{A}})}{\min_{n \neq n'} d(\mathbf{a}_n, \mathbf{a}_{n'})}. \qquad (37)$$

The RMDE is zero if and only if the estimate is perfect, while if the RMDE is close to one, the estimation error is of the same order of magnitude as the distance between true directions,

[2]We thank S. Rickard and C. Fearon for graciously providing their implementation of DUET [2].

indicating a very poor estimation quality. The generalisation of the RMDE for anechoic mixtures is done by replacing in Equations (36) and (37), the distance $d(.,.)$ with the distance $d_c(.,.)$ defined in Equation (29).

C. Evaluation signals

We evaluated the clustering algorithms over the speech data in [27], that is to say speech sources of duration 11.9 s, sampled at 8 kHz from 30 English speakers (males and females) from as many different audio books. For each configuration of the mixing parameters, we generated $T = 10$ mixtures from different source signals.

D. Method parameters

1) STFT parameters: For each method, a STFT $\mathbf{X}(t, f)$ is computed as a first step. Except for DUET, we combine different scales corresponding to frame sizes of 2^n samples, ranging from 128 samples (16 ms) to 65536 samples (8.2 s). For DUET, a frame size of 512 samples (64 ms) is used. We used a Hanning window with a half-frame overlap.

2) DEMIX specific parameters: For DEMIX-Instantaneous and DEMIX-Anechoic, we used the following parameters : the size of the neighborhood is set to $|\Omega| = 10$, which corresponds to the optimum value in the tests of the TIFROM-CF method [16]. The value of the thresholds are set to : $\zeta = 3.3$, $\zeta_c = 9.5$, and (for DEMIX-Anechoic only) $\zeta_2 = 2.33$.

3) ELBG specific parameters: Since ELBG and its variants are randomly initialized, we run them $I = 10$ times for each test mixture and focus on the smallest error (RMDE) over these 10 runs, which thus gives an optimistic estimate of their performance.

E. Evaluations on Instantaneous mixtures

In this section, mixtures are obtained using instantaneous mixtures without noise, that is using Equation (1) with $\delta_{mn} = 0, \forall m, n$ and $n_m(\mathrm{t}) = 0, \forall\, \mathrm{t}$.

1) Experimental protocol: First, we study the performance of the different algorithms depending on the number of sources, and second we fix the number of sources to three, and we vary the distance between these three sources.

a) 1^{st} experiment : N equally spaced sources: in the first experiment, noiseless linear instantaneous mixtures are generated 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 $N = 10$.

b) 2^{nd} experiment : 3 sources getting closer and closer: in the second experiment, 3 sources are placed with the following IP : $\theta_{l+2} = \frac{\pi}{4} + l\,\Delta\theta\,\pi/180$, with $l \in \{-1, 0, 1\}$. In this experiment, we only vary the *angular distance $\Delta\theta$ between IP* in order to test the robustness of the algorithm when directions get close to each other ($\Delta\theta$ small).

2) Results: We observe (Tab I) that up to $N = 8$ sources, DEMIX estimates correctly the number of directions in more than seven cases out of ten, but when $N > 9$ it fails to count the number of sources.

As can be seen on Figures 5 and 6, DEMIX-Instantaneous (DEMIX-Inst) yields a better performance than the best among $I = 10$ instances of the ELBG algorithms. The comparison of DEMIX-Instantaneous with DEMIX-Instantaneous-BR (DEMIX-Inst-BR) shows that the weighted mean using the confidence measure to estimate the directions as opposed to taking the "best" region, significantly improves the performance.

nb of sources	2	3	4	5	6	7	8	9	10
DEMIX Inst	100	100	100	100	100	100	70	20	0

TABLE I
FREQUENCY OF CORRECT COUNT OF THE NUMBER OF SOURCES (IN %)

A remarkable fact is the behavior of DEMIX with three directions getting very close, compared to all other algorithms. The RMDE of the four ELBG variants approaches 1 when the distance between true directions gets close to zero (see Figure 6). In other words, the ELBG variants essentially confuse all directions. On the opposite, *DEMIX remains very robust when the directions are very close to each other*: as reported in Figure 6, the RMDE of DEMIX-Instantaneous (DEMIX-Inst-M) remains below 10^{-3} until the directions get closer than 10^{-3} degrees ($\Delta\theta < 10^{-3}$).

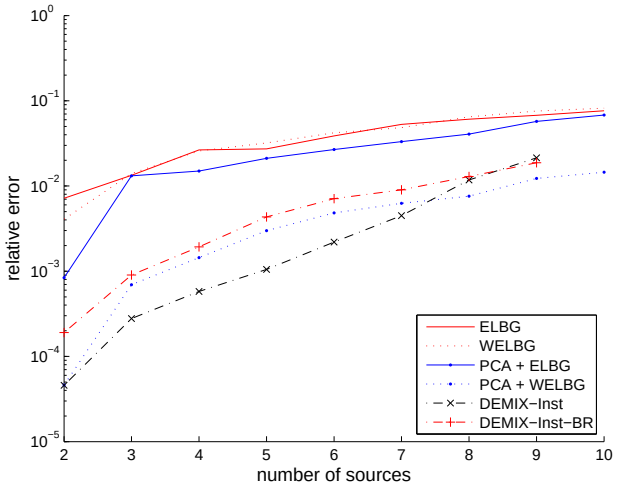


Fig. 5. Relative mean direction error (RMDE) as a function of the number of sources for DEMIX-Instantaneous (DEMIX-Inst), its variant DEMIX-Instantaneous-BR (DEMIX-Inst-BR) and the best instance (over 10) of the four variants of the ELBG

Also we notice, by observing the results for the four variants of the ELBG, that the ELBG results are not significantly improved if the local estimation $\theta(t, f)$ of a direction are replaced by those obtained from local PCA $\widehat{\theta}(t, f)$. The use of the confidence measure to "boost" the most reliable IP has a much more significant impact on the performance.

F. Evaluations on synthetic anechoic mixtures

We propose 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.

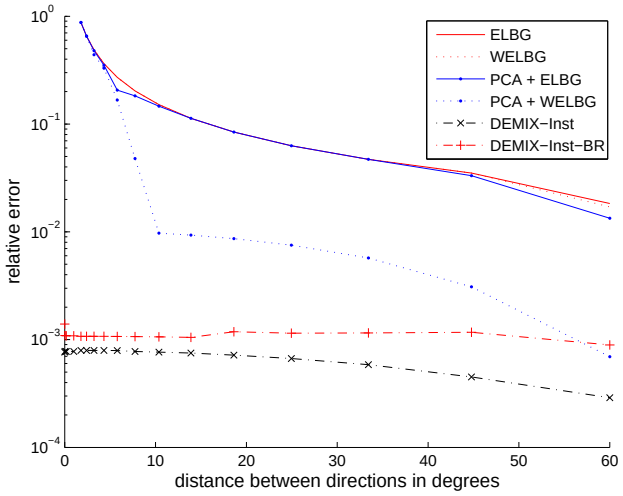
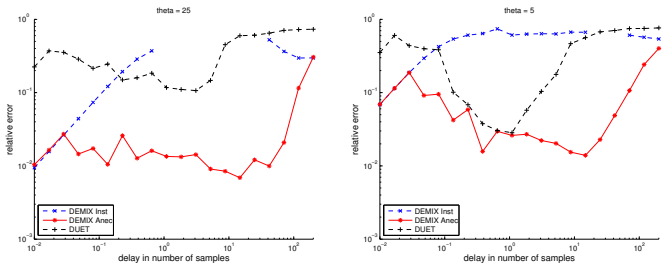


Fig. 6. Relative mean direction error (RMDE) as a function of the angular distance in degrees ($\Delta\theta$) between the 3 directions for DEMIX-Instantaneous (DEMIX-Inst), its variant DEMIX-Instantaneous-BR (DEMIX-Inst-BR) and the best instance (over 10) of the four variants of the ELBG



(a) angular distance between IP is $\Delta\theta = 25$ degrees.

(b) angular distance between IP is $\Delta\theta = 5$ degrees.

Fig. 7. Relative mean direction error (RMDE) as a function of the absolute delay δ of the two side sources.

nb of sources	2	3	4	5	6	7
DEMIX Anec	80	100	100	90	10	0

TABLE II
FREQUENCY OF CORRECT COUNTS OF THE NUMBER OF SOURCES (IN %)

1) Experimental protocol: Similarly to the second experiment in Section VII-E1 we generated stereophonic mixtures of 3 sources with the following IP : $\theta_{l+2} = \frac{\pi}{4} + l\,\Delta\theta\,\pi/180$ with $l \in \{-1; 0; 1\}$ and delays $\delta_n \in \{-\delta, 0, +\delta\}$. The value of the delay $\delta \geq 0$ represents the degree of "anechoism" which is varying. We run experiments with differents values of the angular distance $\Delta\theta$, and report in Figures 7(a) and 7(b) the results for values $\Delta\theta = 25$ and 5 degrees.

2) Results:

a) counting the sources (Figure not reported): As soon as the delay δ is non negligeable, the percentage of correct count for DEMIX-Instantaneous falls below 60%, as opposed to the success of DEMIX-Anechoic. Depending on the value of $\Delta\theta$, this happens when the delay exceeds 0.1 to 1 sample.

b) estimating the directions: For a delay below 0.03 sample, DEMIX-Instantaneous and DEMIX-Anechoic provide similar performance; for larger delays, DEMIX-Instantaneous performance rapidly decreases down to a relative error close to 1.

The larger the angular distance $\Delta\theta$ between IP, the larger the relative error of DUET. This is probably due to the fact that for the clustering step, DUET uses the ratio $R_{21}(\tau,\omega) = X_2(\tau,\omega)/X_1(\tau,\omega)$ which becomes unstable when the IP θ_n get closer to $\pi/2$. Thus, for any delay, the performance of DUET is worse than that of DEMIX-Anechoic. Also, for any angular distance between IP ($\Delta\theta = 5$ or 25 degrees), DEMIX-Anechoic performs better than DUET when the delay is lower than 0.1 sample, but especially when the delay δ is higher than one sample[3].

G. Comparison between DEMIX-Anechoic and DUET on room simulated anechoic mixtures

As a third experiment, we compared the performance 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 [28].

1) Experimental protocol: the experimental protocol is the same as in [8], but applied to the signals described in section VII-C. So we invite the reader to refer to [8] for a detailed description of the protocol.

The experience consists in estimating the performance of algorithms by changing the number of sources from $N = 2$, to $N = 7$.

2) Results:

a) counting the sources: DEMIX-Anechoic estimates the number of sources with more than 80% of success until $N = 5$ (see Tab II).

b) direction estimation: Tab III shows that the average RMDE of DEMIX-Anechoic is consistently better than that of DUET by a factor of at least 10.

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.

nb of sources	2	3	4	5	6
DEMIX Anec	0.005	0.003	0.020	0.018	0.020
DUET	0.242	0.270	0.898	0.449	0.452

TABLE III
AVERAGE RMDE AS A FUNCTION OF THE NUMBER OF SOURCES

H. Evaluation in reverberant situation

In this section, we evaluate the robustness of DEMIX-Anechoic and DUET when there is some reverberation, which corresponds to a more realistic recording situation than the instantaneous and anechoic situation.

1) Experimental protocol: similarly to the experimental protocol of section VII-F, we generated a sterephonic mixture of 3 sources with IP : $\theta_{l+2} = \frac{\pi}{4} + l\,\Delta\theta\,\pi/180$, with $\Delta\theta \in \{5, 25\}$ degrees and with delays $\delta_n \in \{-1, 0, +1\}$, to enable DUET to correctly estimate the delay as follows.

[3]Note that other methods like TIFROM [16] and TIFCORR [17] are able to estimate delays longer than only one sample. Thus, it will be interesting to evaluate all these methods within the framework of an evaluation campaign.

The mixture was obtained as the sum of an anechoic part and a reverberant part :

$$x_m(\mathrm{t}) = \sum_{n=1}^{N} \left(\underbrace{a_{mn} s_n(\mathrm{t} - \delta_{mn})}_{anechoic\ part} + \underbrace{\int b_{mn}(\tau) s_n(\mathrm{t} - \tau) d\tau}_{reverberant\ part} \right) \quad (38)$$

The reverberant part was obtained by generating $b_{mn}(\tau)$ as an independant Gaussian noise $b_{mn}(\tau) \sim \mathcal{N}\left(0, \sigma^2\left(\tau - \delta_{mn} - t_0\right)\right)$ with :

$$\sigma^2(\tau) = \begin{cases} 10^{-\alpha\tau}\sigma_N^2 & \text{if } 0 \leq \tau < K \\ 0 & \text{otherwise} \end{cases},$$

with $\alpha = 6/K$, so as to have an exponential decrease of -60 dB at the end of the reverberation part, and with $t_0 = 50$ ms, and $K = 150$ ms. The parameter σ_N^2 controls the input SNR defined as : $SNR_{in} = 10\log_{10}\left(\frac{\sum_{m,n} a_{mn}^2}{\sum_{m,n}\int b_{mn}^2(\tau)d\tau}\right)$.

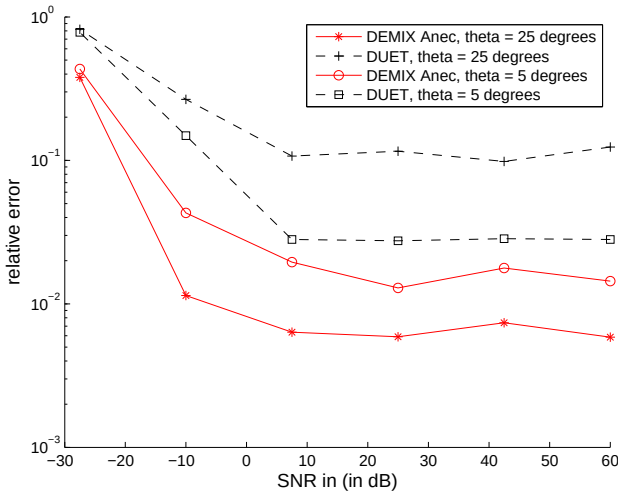


Fig. 8. Relative mean direction error (RMDE) between the estimated anechoic mixture directions and the anechoic part of the true mixture directions, as a function of the SNR_{in}, for values $\Delta\theta = 25$ degrees and $\Delta\theta = 5$ degrees of the angular distance between IP, and for an absolute delay $\delta = 1$ of the two side sources.

2) Results: Figure 8 shows that the RMDE between the estimated anechoic mixture directions and the anechoic part of the true mixture directions decreases when the SNR_{in} gets lower than 0 dB, and the lower the SNR_{in}, the higher the RMDE for the two methods. This is due to the fact that the convolutive model of Equation (38) differs more and more from the anechoic one of Equation (1) but probably also because adding reverberation decreases the sparsity of the source images in the TF domain. Also, whatever is the SNR_{in} DEMIX-Anechoic obtains better results than DUET, whereas the absolute delay δ of the two side sources is set to 1, which is the most favorable configuration for the DUET method according to Figures 7(a) and 7(b).

VIII. Summary and Conclusions

We have 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. The experiments on stereophonic recordings have illustrated the ability of the proposed method : a) to count the number of sources until 8 sources in the instantaneous case and 5 sources in the anechoic case; 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 100 samples in simulated anechoic mixtures.

The proposed method relies on a 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 as well as the number of sources, 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. This condition is likely to fail when the mixture is made of "too many" sources or when the sources representations are not sparse enough. 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 found to cluster directions estimated at different frequencies.

References

[1] P. Bofill and M. Zibulevsky, "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] E. Vincent, "Complex nonconvex l_p norm minimisation for underdetermined source separation," in *ICA'07*. Springer, 2007, pp. 430–437.

[4] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *ICA'09*.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. with application to blind audio source separation," in *ICASSP'09*.

[6] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *ICA'09*.

[7] 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.

[8] ——, "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.

[9] A. Belouchrani and M. Amin, "Blind source separation based on time-frequency signal representations," *Signal Processing, IEEE Transactions on*, vol. 46, no. 11, pp. 2888–2897, Nov 1998.

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

[11] 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.

[12] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," *EURASIP Journal on Applied Signal Processing*, vol. 17, p. 2828, 2005.

[13] C. 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.

[14] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *ICA*, 2009, pp. 734–741.

[15] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, Issues 3-4, no. 3, pp. 627–641(15), October 2003.

[16] 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.

[17] ——, "A time-frequency correlation-based blind source separation method for time-delayed mixtures," in *ICASSP*, 2006.

[18] Y. Luo and J. Chambers, "Active source selection using gap statistics for underdetermined blind source separation," in *Proc. 7th Int. Symp. on Signal Processing and Its Applications (ISSPA)*, vol. 1, 2003.

[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis, 2nd Edition*. Wiley-Interscience, September 1984.

[20] ——, "Asymptotic theory for principal component analysis," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963.

[21] R. Duda and P. Hart, *Pattern classification ans Scene Analysis*. New York: John Wiley, 1973.

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

[23] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *Communications, IEEE Transactions on*, vol. 28, no. 1, pp. 84–95, Jan 1980.

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

[25] R. Xu and D. W. II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

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

[27] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms." *Signal Processing*, 2007.

[28] D. Campbell, K. Palomäki, 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.

Simon Arberet was born in France in 1980. He graduated from École Supérieure d'Électricité (Supélec), Paris, France in 2003. He received the BA degree in music from University of Paris 8, France, in 2004 and the M.S. degree in computer science from the Pierre and Marie Curie University (Paris 6), France in 2005. He received the Ph.D degree in signal processing from Univesity of Rennes 1, France in 2008. He was with the National Center for Scientific Research (CNRS) at IRISA, Rennes, France from 2005 to 2008 while working towards his Ph.D degree. He is currently working with the National Institute for Research in Computer and Control Science (INRIA) at IRISA, Rennes, France. His research interests include statistical signal processing, sparse signal representation and blind audio source separation.

Rémi Gribonval graduated from École Normale Supérieure, Paris, France in 1997. He received the Ph. D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, France, in 1999, and his Habilitation à Diriger des Recherches in applied mathematics from the University of Rennes I, Rennes, France, in 2007. He is a Senior Member of the IEEE.

From 1999 until 2001 he was a visiting scholar at the Industrial Mathematics Institute (IMI) in the Department of Mathematics, University of South Carolina, SC. He is now a Senior Research Scientist (Directeur de Recherche) with INRIA (the French National Center for Computer Science and Control) at IRISA, Rennes, France, in the METISS group. His research focuses on sparse approximation, mathematical signal processing and applications to multichannel audio signal processing, with a particular emphasis in blind audio source separation and compressed sensing. Since 2002 he has been the coordinator of several national, bilateral and european research projects, and in 2008 he was elected a member of the steering committee for the international conference ICA on independent component analysis and signal separation.

Frédéric Bimbot received the B.A. degree in linguistics from Sorbonne Nouvelle University, Paris, France, in 1987, the telecommunication engineer degree from ENST, Paris, in 1985, and the Ph.D. degree in signal processing in 1988. In 1990, he joined the French National Center for Scientific Research (CNRS) as a Permanent Researcher. He was with ENST for seven years and then moved to IRISA (CNRS and INRIA), Rennes, France. He also repeatedly visited AT&T Bell Laboratories between 1990 and 1999. He has been involved in several European projects: SPRINT (speech recognition using neural networks), SAM-A (assessment methodology), and DiVAN (audio indexing). He has also been the work-package Manager of research activities on speaker verification, in the projects CAVE, PICASSO, and BANCA. His research is focused on audio signal analysis, speech modeling, speaker characterization and verification, speech system assessment methodology, and audio source separation. He is heading the METISS Research Group at IRISA, dedicated to selected topics in speech and audio processing. Dr. Bimbot was Chairman of the Groupe Francophone de la Communication Parle (now AFCP) from 1996 to 2000 and from 1998 to 2003, a member of the International Speech Communication Association Board (ISCA), formerly known as ESCA.