A ROBUST METHOD TO COUNT AND LOCATE AUDIO SOURCES IN A STEREOPHONIC LINEAR ANECHOIC MIXTURE

Simon Arberet, Remi Gribonval, Frédéric Bimbot

IRISA (INRIA & CNRS), Rennes France

ABSTRACT

We propose a new method, called DEMIX Anechoic, to estimate the mixing conditions, i.e. number of audio sources plus attenuation and time delay of each sources, in an underdetermined anechoic mixture. The method relies on the assumption that in the neighborhood of some time-frequency points, only one source contributes to the mixture. Such time-frequency points, located with a local confidence measure, provide estimates of the attenuation, as well as the phase difference at some frequency, of the corresponding source. The time delay parameters are estimated, by a method similar to GCC-PHAT, on points having close attenuations. As opposed to DUET like methods, our method can estimate time-delay higher than only one sample. Experiments show that DEMIX Anechoic estimates, in more than 65% of the cases, the number of directions until 6 sources and outperforms DUET in the accuracy of the estimation by a factor of 10.

Index Terms— Signal analysis, Discrete Fourier transforms, Delay estimation, Audio recording, Cognitive science

1. INTRODUCTION

The problem of estimating the number of audio sources and the mixing directions is considered in a possibly degenerate linear anechoic mixture. In the time domain, we have :

$$x_m(\tau) = \sum_{n=1}^{N} a_{mn} s_n(\tau - \delta_{mn}), \qquad m = 1, 2, \ldots, M$$

where $M \leq N$, $a_{mn} \in \mathbb{R}^+$ and $\delta_{mn} \in \mathbb{R}$ are attenuation coefficients and time delays associated with the path from n^{th} source to the m^{th} microphone. Whithout loss of generality, we set $\delta_{1n} = 0$ for $n = 1, 2, \ldots, N$, and $\sum_{m=1}^{M} a_{mn}^2 = 1$.

Taking the Short Time Fourier Transform (STFT) of x_1,-$x_2, \ldots, x_M$, the mixing model can be written in a matrix form $\hat{\mathbf{x}}(t,f) = \mathbf{A}(f)\hat{\mathbf{s}}(t,f)$ with $\hat{\mathbf{x}} = [\hat{x}_1 \ldots \hat{x}_M]^T$, $\hat{\mathbf{s}} = [\hat{s}_1 \ldots \hat{s}_N]^T$. In the stereophonic case, i.e $M = 2$, the mixing matrix is :

$$\mathbf{A}(f) = \begin{bmatrix} a_{11} & \ldots & a_{1N} \\ a_{21}e^{-i2\pi f \delta_1} & \ldots & a_{2N}e^{-i2\pi f \delta_N} \end{bmatrix}$$

We replaced the δ_{2n} notation with δ_n for simplificity. As each column of $\mathbf{A}(f)$ is normalised, a source direction n is defined by only two parameters :

1. The intensity difference (ID) θ_n

$$\theta_n := \tan^{-1}(a_{2n}/a_{1n}) \tag{1}$$

2. The delay δ_n

Several methods exist that attempt to estimate the mixing directions, that is to say δ_n and θ_n. DUET [1] and TIFROM [2] are based on a time-frequency representation of the observed signals and exploit the fact that at some time-frequency points, only one source contributes to the mixture. This assumption is related to sparsity of the time-frequency representation of the sources. Our approach relies on the same assumption.

The DUET method finds directions by estimating the ID and delay parameters directly on each time-frequency point, and by finding maxima in a smoothed histogram of these parameters. However the drawback of the DUET method, is that it cannot estimate delays higher than one sample. This problem has already been reported in [1, 3], and a solution was proposed by Puigt [3] but with few experimental results. Notice that a delay of one audio sample can correspond to a very short distance : At a CD sampling rate of 44.1kHz, a delay of one sample correspond physically to a distance of propagation in the air of 7.8 mm.

In addition to many other source-separation approaches which exploit globally the sparsity property [4, 1], our approach exploits local estimates of the activity/inactivity of each source to get a more robust estimation of the ID and delay parameters. This local approach, which has already been used in the TIFROM method [2], is less sensitive to the sparsity assumption.

Our main contribution is to propose a new clustering algorithm called DEMIX Anechoic, which extends the DEMIX instantaneous algorithm [5] to the anechoic case. To do so we : (1) introduce a confidence measure to determine how valid is the assumption that only one source contributes to the mixture in a given time-frequency region, and a way to estimate parameters of this time-frequency region by the use of a "complex" Principal Component Analysis (PCA) ; (2) propose a new way to estimate delays without the restriction of having delays lower than only one sample. For this we introduce a weighted correlation function similar to the GCC-PHAT method [6].

Section 2 presents our approach, section 3 details the DEMIX Anechoic algorithm we propose, and section 4 shows the performances of our algorithm compared with other ones.

2. PROPOSED APPROACH

2.1. Exloiting sparsity and consistence

Sparsity enables to easily identify the ID parameter θ_n (1). Let suppose we have a sparse source model where only one source $n := n(t,f)$ is active in each time-frequency point. That is $\hat{s}_n(t,f) \neq 0$ and $\hat{s}_k(t,f) = 0\ \forall k \neq n$. In such case $\mathbf{x}(t,f) = \mathbf{a_n}(f)s_n(t,f)$, where $\mathbf{a_n}(f) = \left[a_{1n}\ a_{2n}e^{-i2\pi f\delta_n}\right]^T$ is the direction n. In this ideal case, we can easily estimate the ID parameter by taking the DUET ratio $R_{21} := \frac{\hat{x}_2}{\hat{x}_1}$ [1]. Indeed, if only source n is active, $R_{21}(t,f) \approx \frac{a_{2n}}{a_{1n}}e^{-i2\pi f\delta_n}$. So, by taking the absolute value $|R_{21}(t,f)|$, we obtain the ID parameter of direction n, that is $\theta_n = \tan^{-1}|R_{21}(t,f)|$.

In DUET [1], the delay is estimated with the phase of the R_{21} ratio : $\tilde{\delta}(t,f) := -\frac{1}{2\pi f}\angle R_{21}(t,f)$. However this estimation is ambiguous if the delay is higher than one sample. Suppose that only source n is active. If we assume that the delay δ_n is less than one sample, as $f < \frac{1}{2}$, we have : $2\pi\delta_n f + 2k\pi \in [-\pi\ \pi]$ if and only if $k = 0$. So the phase of R_{21} is not ambiguous modulo $2k\pi$. But if we don't assume that the delay δ_n is less than one sample, we cannot deduce k, and the phase ambiguity is not resolvable. So the delay estimation : $\tilde{\delta}(t,f) \approx \delta_n + \frac{k}{f}$ is biased with an unknown value $\frac{k}{f}$. In other words, as several δ_n are compatible with the phase $\angle R_{21}(t,f)$, a time-frequency point has not enough information to deduce the delay of its direction. For this reason, it is necessary to gather several time-frequency points of the same source at different frequencies. That raises two issues. (1) How to find several points of the same source ? (2) How to deduce the delay, if we suppose that we have several points of the same source ? A new approach to estimate the delay without this ambiguity problem is proposed in section 2.3.

However in some time-frequency points, several sources are simultaneously active. In this case, it is difficult to estimate the mixing directions by simply clustering all the points, because some sources of weak energy may not appear clearly. Our approach, inspired by TIFROM [2] and already used for DEMIX Instantaneous [5], consists in "boosting" points that have a great chance of being generated by only one source. To do this, we first define time neighborhoods $\Omega_{t,f} = \{(t + kL/2, f) | |k| \leq K\}$ around each time-frequency point $\mathbf{x}(t,f)$. K is the neighborhood size and L is the STFT window size. In order to account for the different possible durations of audio structures in each source, we use a multi-resolution framework, so L has different values. If we assume that only source n is active in the neighborhood $\Omega_{t,f}$, the points $\mathbf{x}(t,f)$ are all aligned with the same "complex" direction $\mathbf{a_n}(f)$ because $\mathbf{x}(t,f) = \mathbf{a_n}(f)s_n(t,f)$. Whereas if we assume that several sources are simultaneously active in $\Omega_{t,f}$, points are no longer aligned.

So, by detecting how strongly the points of $\Omega_{t,f}$ are aligned in the principal direction, we have a measure (see section 2.2) which shows if only one source is present or not. That is a measure which shows if the estimated direction points or not to a direction of the mixing matrix at frequency f.

To get this principal direction, we compute a Principal Component Analysis (PCA) on the time-frequency points of the neighborhood. In other words, we extract the eigenvector of the highest eigenvalue of the $2 \times (2K+1)$ matrix $\mathbf{X}_{\Omega_{t,f}}$ with entries $\mathbf{x}(t,f)$. We obtain a principal direction $\hat{\mathbf{u}}(t,f) = [u_1(t,f)\ u_2(t,f)]^T \in \mathbb{C}^2$ which is translated as follows : $\widehat{\theta}(t,f) = \tan^{-1}\left(\left|\frac{u_2(t,f)}{u_1(t,f)}\right|\right)$, $\widehat{\phi}(t,f) = \angle\left(\frac{\hat{u}_2(t,f)}{\hat{u}_1(t,f)}\right)$.

2.2. The confidence measure

To have an idea of how likely it is that the unit principal vector $\hat{\mathbf{u}}(t,f)$ of the PCA on $\Omega_{t,f}$ corresponds to the direction of the most active source $\mathbf{a_n}(f)$ at frequency f, we need to know with what confidence we can trust the fact that a single source is active in $\Omega_{t,f}$.

For that, we model the STFT coefficients of the most active source s in a neighborhood $\Omega_{t,f}$, as well as the contribution of all other sources plus possibly noise $\mathbf{n}$, with centered circular normal distributions. That is $s \sim \mathcal{N}_c\left(0, \sigma_s^2\right)$, and $\mathbf{n} \sim \mathcal{N}_c\left(0, \sigma_n^2\mathbf{I}_2\right)$. The model for points $(t',f') \in \Omega_{t,f}$ is :

$$\mathbf{x}(t',f') = \mathbf{a}_n(f)s(t',f') + \mathbf{n}(t',f')$$

So, $\mathbf{x}(t',f') \sim \mathcal{N}_c\left(0, \Sigma_{\mathbf{x}}\right)$, with $\Sigma_{\mathbf{x}} = \sigma_n^2\mathbf{I}_2 + \sigma_s^2\mathbf{a}_n(f)\mathbf{a}_n^H(f)$. We define the confidence measure as $\mathcal{T} := \frac{\lambda_1}{\lambda_2}$, where $\lambda_1 \geq \lambda_2$ are the eigenvalues of $\Sigma_{\mathbf{x}}$. We show that if $\lambda_1 > \lambda_2$, $\mathbf{a}_n(f)$ is the eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to λ_1, and that $\mathcal{T} = 1 + \frac{\sigma_s^2}{\sigma_n^2}$. So the confidence measure can be viewed as a signal to noise ratio between the dominant source and the contribution of the other ones plus noise.

As $\hat{\mathbf{u}}(t,f)$ is computed by PCA on sample of $m := \text{card}(\Omega_{t,f})$ points, it only provides an estimate of the direction $\mathbf{a}_n(f)$, with a precision we want to estimate. This precision is defined by equation (2).

$$d^2(\hat{\mathbf{u}}(t,f), \mathbf{a}_n(f)) := 2\left(1 - |\langle\hat{\mathbf{u}}(t,f), \mathbf{a}_n(f)\rangle|\right) \quad (2)$$

For a large sample size, we show that $d(\hat{\mathbf{u}}(t,f), \mathbf{a}_n(f))$ converges in law to $\mathcal{N}\left(0, \sigma^2(\mathcal{T})\right)$ with

$$\sigma^2(\mathcal{T}) := \frac{1}{m-1}\frac{\mathcal{T}}{(\mathcal{T}-1)^2} \quad (3)$$

In the Time-Delay estimation, and in the clustering algorithm we present in next sections, we use the variance of equation (3). However, as we don't know the confidence measure $\mathcal{T}$, we use the empirical confidence measure $\widehat{\mathcal{T}}$ computed by PCA instead. This measure is defined by : $\widehat{\mathcal{T}} := \frac{\hat{\lambda}_1}{\hat{\lambda}_2}$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2$ are the eigenvalues of $\mathbf{X}_{\Omega_{t,f}}$.

2.3. Time-Delay estimation

Suppose that only one source n is active If we compute the (Inverse Fourier Transformation) IFT on a frame t of the phase part of the estimated point, i.e. $e^{i\widehat{\phi}(t,f)} \approx e^{-i2\pi f\delta_n}$, we get a Dirac on the delay δ_n. Unfortunately, we are not in this ideal case and several sources are active. So we select a set of points which have a great chance to belong to the same source, by selecting points which have an ID close to the ID of a point having a high confidence. A weighted sum of $e^{i\widehat{\phi}(t,f)}$ over the frames is then computed, using the confidence measure as weight (the goal is to favor points where only one source is active). Then we get the highest peak of the IFT as delta estimate : $\widehat{\delta_k} = arg\max_{\delta_n} \widehat{\mathcal{R}}(\delta_n)$ with

$$\widehat{\mathcal{R}}(\delta_n) := \int \frac{1}{\sum_t \sigma^{-2}(\widehat{\mathcal{T}}(t,f))} \sum_t \frac{e^{i\widehat{\phi}(t,f)} e^{i2\pi f\delta_n}}{\sigma^2(\widehat{\mathcal{T}}(t,f))} df \quad (4)$$

$\sigma^2(.)$ is defined in equation (3). If other directions with a similar ID are present, the highest peak will correspond to one of these directions. By removing points near the direction that have the highest peak, the other directions can be estimated in a later iteration. Notice that this delay estimator is a variation of the GCC-PHAT estimator [6].

3. DEMIX ANECHOIC ALGORITHM

The first step of the algorithm consists in iteratively creating K clusters by : (1) selecting points $(\widehat{\theta}_k, \widehat{\phi}_k, \widehat{\mathcal{T}}_k)$ with highest confidence, (2) estimating the delay $\widehat{\delta}_k$ corresponding to the cluster with points having $\widehat{\theta}$ closed to $\widehat{\theta}_k$, (3) creating the cluster by aggregating points near the centroid $(\widehat{\theta}_k, \widehat{\delta}_k)$. The second step is to re-estimate the direction $(\widehat{\theta}_k, \widehat{\delta}_k)$ of each cluster, and finally to eliminate non significant clusters and keep $\widehat{N} \leq K$ clusters which centroids provide the estimated directions of the mixing matrix.

3.1. Cluster Creation and Delay Estimation

DEMIX iteratively creates K clusters $C_k \subset P$ –where P is the set of all points– starting from $K = 0$, $P_K = P_0 = P$:

1. find the point $(\widehat{\theta}_K, \widehat{\phi}_K, \widehat{\mathcal{T}}_K) \in P_K$ with the highest confidence ;
2. create a temporal cluster $\widetilde{C}_K$ with all points $(\widehat{\theta}, \widehat{\phi}, \widehat{\mathcal{T}}) \in P_K$ which have their $\widehat{\theta}$ "sufficiently close" to $(\widehat{\theta}_K, \widehat{\mathcal{T}}_K)$;
3. estimate $\widehat{\delta}_K$ with our Time-Delay estimation method applied to $\widetilde{C}_K$ points ;
4. – if there is not a "well identified" $\widehat{\delta}_K$: reject the cluster, $P_{K+1} := P_K \setminus \widetilde{C}_K$;
   – else, create the cluster C_K with all points $(\widehat{\theta}, \widehat{\phi}, \widehat{\mathcal{T}}) \in P_K$ "sufficiently close" to $(\widehat{\theta}_K, \widehat{\delta}_K, \widehat{\mathcal{T}}_K)$, and $P_{K+1} := P_K \setminus C_K$;
5. if $P_{K+1} = \emptyset$, stop ; otherwise increment $K \leftarrow K + 1$ and go back to 1.

Expressions "sufficiently close" rely on the model developped in section 2.2. Expression "sufficiently close" to $(\widehat{\theta}_K, \widehat{\mathcal{T}}_K)$ in step 2, includes all points $(\widehat{\theta}, \widehat{\phi}, \widehat{\mathcal{T}}) \in P_K$ such that $\left|\widehat{\theta} - \widehat{\theta}_K\right| \leq \sigma_1(\widehat{\mathcal{T}}_K)$, where the expression of $\sigma_1(\widehat{\mathcal{T}}_K)$ will be detailed in a futur paper. Expression "sufficiently close" to $(\widehat{\theta}_K, \widehat{\delta}_K, \widehat{\mathcal{T}}_K)$ in step 4 includes all points $(\widehat{\theta}, \widehat{\phi}, \widehat{\mathcal{T}}) \in P_K$ such that $\frac{1}{\Delta_f} \int d(\hat{\mathbf{u}}, \hat{\mathbf{u}}_K) df \leq \sigma_2(\widehat{\mathcal{T}}, \widehat{\mathcal{T}}_K)$ where Δ_f is the frequency domain, d is the distance defined in (2), $\hat{\mathbf{u}} = [\cos(\widehat{\theta}) \ \sin(\widehat{\theta}) e^{i\widehat{\phi}}]^T$, $\hat{\mathbf{u}}_K = [\cos(\widehat{\theta}_K) \ \sin(\widehat{\theta}_K) e^{-i2\pi\widehat{\delta}_K f}]^T$, and $\sigma_2(\widehat{\mathcal{T}}, \widehat{\mathcal{T}}_K)$ is defined in equation (8) of paper [5].

Expression there is a "well identified" $\widehat{\delta}_K$ in step 4 means that, no other peaks higher than -3dB of $\widehat{\delta}_K$ appears in the Time-Delay estimation function.

3.2. Direction Estimation and Cluster Elimination

The direction re-estimation, and the cluster elimination steps are similar to the 2^d and 3^d step of the first version of DEMIX Instantaneous [5]. The main difference is that a new measure of centroid distance based on equation (2) is used to consider the phase difference induced by each direction which changes with frequency.

4. EXPERIMENTS

4.1. Experimental protocol

We compare ability of DEMIX Anechoic, DEMIX Instantaneous, and DUET, to estimate the directions of some anechoic mixtures. The RoomSim MATLAB simulation of room was used in order to generate anechoic mixing matrices. Two cardiod microphones were placed at 20 cm from each other, and their directions crossed with a right angle. Sources were placed on 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 1).

The source selection process was the same as in the DEMIX Instantaneous paper [5] (polish voices sampled at 4kHz). The experience consisted in estimating the performance of algorithms by changing the number of sources from $N = 2$, to $N = 7$.

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 better than DEMIX Instantaneous (see figure 3). However DEMIX Anechoic always fails when $N > 6$. Note that we cannot compare these results with DUET, because DUET doesn't estimate the number of sources and takes it as an input.

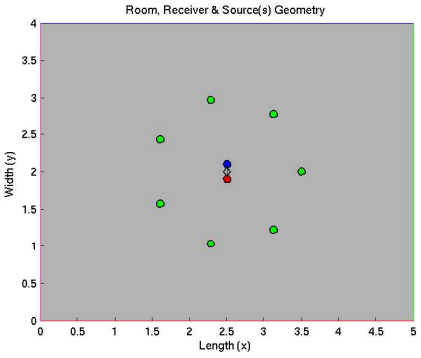

Fig. 1. room configuration for $N = 7$ sources surrounding the stereo microphone pair

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

Fig. 2. Tab shows the θ (in radians) and δ (in samples) parameters corresponding to the room configuration of figure 1

In case of success, we could also measure the means over test mixtures of the *direction distance mean error* (DDME), which is the mean distance between true directions and estimated ones.
$DDME(\mathbf{U}, \mathbf{A}) = \frac{1}{N} \sum_{i=1}^{N} \int_f d(\widehat{\mathbf{U}}_i(f), \mathbf{A}_i(f))df$, where $\widehat{\mathbf{U}_i}(f) = [\cos(\widehat{\theta_i}) \ \sin(\widehat{\theta_i})e^{-i2\pi\widehat{\delta}_i f}]^T$ is the estimated direction corresponding to the direction $\mathbf{A}_i(f)$. Figure 4 shows that DEMIX Anechoic obtained a lower DDME error than DEMIX Instantaneous and than DUET. DUET worked with a weighted K-Means algorithm as implemented by its authors [1]. Since the DDME for DEMIX can only be measured when a correct number of sources is estimated, it was not computed when $N > 6$ with DEMIX Anechoic, and for $N = 2$ and $N > 5$ with DEMIX Instantaneous. In any cases DDME for DUET was computed with the same test mixtures as these used with DEMIX Anechoic.

Two major reasons can explain why DEMIX Anechoic obtained best result as DUET. First delays that are higher than one sample result in ambiguous delay estimations for DUET contary to DEMIX Anechoic. Second directions that are near the poles (θ is near 0 or $\pi/2$) are badly estimated by DUET because their ID parameter is *asymmetric* and especially inaccurates for directions located near the poles.

nb of sources	2	3	4	5	6	7
DEMIX Inst	0	65	30	35	0	0
DEMIX Anec	90	100	95	65	5	0

Fig. 3. good number of sources estimation ratio (in %)

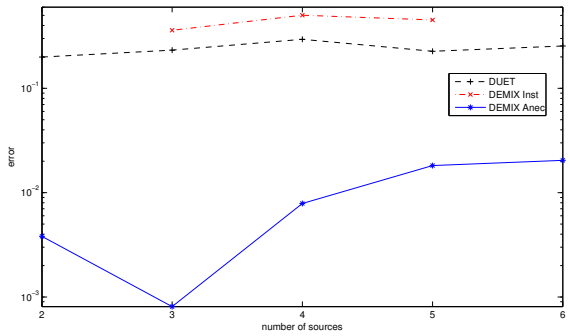

Fig. 4. average DDME as a function of the number of sources

5. CONCLUSION

We have presented a new algorithm called DEMIX Anechoic, to estimate the number of sources, and the mixing directions for under-determined anechoic mixtures. DEMIX Anechoic exploits locally the sparsity of the time-frequency representation, and extracts the parameters of the mixing model via clustering, using a confidence measure. The confidence measure allows to reliably detect regions of time-frequency points where essentially one source is active. As opposed to DUET, DEMIX Anechoic estimates by itself the number of sources, and can estimate delays that are higher than one sample. Moreover, by considering each pair of microphones, the problem originally designed for two microphones, can be extended to M microphones.

6. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE Transactions on Signal Processing*, July 2004 2002, vol. 52, pp. 1830–1847.

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

[3] M. Puigt and Y. Deville, "A time-frequency correlation-based blind source separation method for time-delayed mixtures," in *ICASSP*, 2006.

[4] M. Zibulevsky P. Bofill, "Underdetermined blind source separation using sparse representations," in *Signal Processing*, 2001, vol. 81, pp. 2353–2362.

[5] 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.

[6] C. H. Knapp and G.C. 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.