



**HAL**  
open science

## Multi-source TDOA estimation in reverberant audio using angular spectra and clustering

Charles Blandin, Alexey Ozerov, Emmanuel Vincent

► **To cite this version:**

Charles Blandin, Alexey Ozerov, Emmanuel Vincent. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. [Research Report] RR-7566, 2011, pp.22. inria-00576297v1

**HAL Id: inria-00576297**

**<https://inria.hal.science/inria-00576297v1>**

Submitted on 14 Mar 2011 (v1), last revised 11 Oct 2011 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

***Multi-source TDOA estimation in reverberant audio  
using angular spectra and clustering***

Charles Blandin — Alexey Ozerov — Emmanuel Vincent

**N° 7566**

Mars 2011

— Audio, Speech, and Language Processing —



***R*** *apport  
de recherche*



## Multi-source TDOA estimation in reverberant audio using angular spectra and clustering

Charles Blandin\*, Alexey Ozerov<sup>†</sup>, Emmanuel Vincent<sup>‡</sup>

Theme : Audio, Speech, and Language Processing  
Perception, Cognition, Interaction  
Équipes-Projets METISS

Rapport de recherche n° 7566 — Mars 2011 — 22 pages

**Abstract:** In this article, we consider the problem of estimating the time differences of arrival (TDOAs) of multiple sources from two-channel reverberant audio mixtures. This is commonly achieved using clustering or angular spectrum-based methods. These methods are limited in that they typically affect the same weight to the spatial information provided by all time-frequency bins and rely on a binary activation model of the sources. Moreover, few experimental comparisons of different methods have been carried out so far. We introduce two new groups of TDOA estimation methods. First, we propose a time-frequency weighting procedure based on a form of signal-to-noise-ratio (SNR) that was shown to be efficient for instantaneous mixtures. Second, we introduce new clustering algorithms based on the assumption that all sources can be active in each time-frequency bin. We also study a two-step procedure combining angular spectra and clustering and conduct a large-scale experimental evaluation of the proposed and existing methods. The best average localization performance is achieved by a variant of the generalized cross-correlation with phase transform (GCC-PHAT) method without subsequent clustering. Moreover, one of the SNR-based methods we propose outperforms this method for small microphone spacing.

**Key-words:** Multiple source localization, TDOA estimation, angular spectrum, clustering

This work was supported in part by the ECHANGE project, funded by ANR, and by the Quaero Programme, funded by OSEO.

\* charles.blandin@inria.fr

† alexey.ozerov@inria.fr

‡ emmanuel.vincent@irisa.fr

## Estimation des TDOAs de plusieurs sources en environnement réverbérant par spectre angulaire et clustering

**Résumé :** Dans cet article, nous considérons le problème d'estimation des différences de temps d'arrivée (TDOAs) de plusieurs sources sonores dans un enregistrement stéréophonique en environnement réverbérant. Ce problème est communément traité par des méthodes de type clustering ou spectre angulaire. Ces méthodes sont limitées par le fait qu'elle affectent typiquement le même poids à l'information spatiale issue de tous les points temps-fréquence et qu'elles se basent sur un modèle binaire d'activation des sources. De plus, peu de comparaisons expérimentales ont été effectuées jusqu'à présent. Premièrement, nous proposons une procédure de pondération temps-fréquence basée sur une forme de rapport signal-à-bruit (RSB) dont l'efficacité a été montrée pour des mélanges instantanés. Deuxièmement, nous introduisons de nouveaux algorithmes de clustering basés sur l'hypothèse que toutes les sources peuvent être actives en chaque point temps-fréquence. Nous étudions également une procédure en deux étapes combinant le spectre angulaire et le clustering et nous menons une évaluation expérimentale à grande échelle des méthodes proposées et existantes. En moyenne, les meilleures performances de localisation ont été obtenues par une version de GCC-PHAT (Generalized Cross Correlation with Phase Transform) sans avoir recours au clustering. De plus, une des méthodes basées sur le RSB que nous proposons se révèle plus performante que cette dernière lorsque la distance entre les microphones est petite.

**Mots-clés :** Localisation de sources, estimation des TDOAs, spectre angulaire, clustering

## 1 Introduction

Recorded audio signals often result from a mixture of several sound sources. The problem of source localization consists in estimating the spatial positions of the sources and has many applications such as video-conferencing, surveillance, or source separation [1, 2]. When the signal is recorded by multiple sensors, this problem is often addressed by estimating the *Time Difference Of Arrival* (TDOA) of each source for each pair of sensors [3]. In the following, we focus on TDOA estimation of two or more sources for a given pair of sensors in a reverberant environment.

Most source localization methods use the *Short Time Fourier Transform* (STFT) of the signal [3, 4, 5]. Let us denote by  $\mathbf{x}(t, f) = [x_1(t, f), x_2(t, f)]^T$  and  $s_n(t, f)$ ,  $n = 1, \dots, N$ , the STFTs of the observed signals and the  $n$ -th source signal, where  $t = 1, \dots, T$  and  $f = 1, \dots, F$  are, respectively, time frame and frequency bin indices. With these notations, the source mixing process can be modeled as [6]

$$\mathbf{x}(t, f) = \sum_{n=1}^N \mathbf{d}(f, \tau_n) s_n(t, f) + \mathbf{b}(t, f), \quad (1)$$

where

$$\mathbf{d}(f, \tau_n) = [1, e^{-2i\pi f \tau_n}]^T \quad (2)$$

is the so-called *steering vector* associated with the  $n$ -th source of TDOA  $\tau_n$  (in seconds), and  $\mathbf{b}(t, f)$  models the reverberant part of the signal and additive noise, if any.

We distinguish three general approaches for TDOA estimation. The simplest one consists in computing the TDOA of the mixture signal locally in each time-frequency bin and localizing the peaks of the resulting *histogram* [7, 8, 9]. This approach is restricted to closely spaced or binaural microphones, since in the case of widely spaced microphones local TDOA computation becomes ambiguous due to spatial aliasing [10]. The second approach consists in iteratively estimating the time-frequency bins associated to each source and the corresponding TDOAs by means of some *clustering* algorithm [10, 6, 11]. This approach can be used for any microphone spacing but is sensitive to the initialization of the parameters (i.e., clusters and TDOAs). The third approach [3, 12, 5, 13, 14] consists of building for each time-frequency bin a function of TDOA that is likely to exhibit a large value for true TDOAs, and pooling it across the time-frequency plane so as to obtain a so-called *angular spectrum*. The source TDOAs are then estimated as the highest peaks of this angular spectrum. This approach works for any microphone spacing and does not need any initial guess of the TDOAs. However, the possible presence of secondary peaks can alter the estimation of the TDOAs. In this paper, we focus on the two latter approaches, which are always applicable.

We have identified the following gaps in the state-of-the-art. First, existing methods typically affect the same weight to all time-frequency bins, regardless of the fact that the associated spatial information is less accurate in the presence of several active sources or reverberation. Second, existing clustering-based methods mainly rely on a binary activation model of the sources, which becomes inaccurate when the sources overlap in the time-frequency plane. Third, there

has been no study about the importance of initialization for clustering-based methods. Finally, in all experimental studies, the comparison was made over a small number of algorithms and configurations [15, 16] and few details were given about the localization performance with respect to the microphone spacing, the reverberation time, or the number of sources.

The purpose of this work is to fill in the above-mentioned gaps. First, we define new angular spectra based on a form of *signal-to-noise ratio* (SNR), which affects more weight to the time-frequency bins resulting from a single source. Indeed, the SNR and variants thereof were shown to be efficient confidence measures in the context of clustering-based localization for instantaneous mixtures [17, 18, 19], but have not been extended to angular spectrum-based methods in the context of convolutive mixtures. We consider several time-frequency pooling functions, thereby extending our preliminary results in [20] for a single pooling function. Second, we propose two new clustering-based algorithms. In the first one, both the active source and the noise are modeled as Gaussian random variables under a diffuse noise [6, 20] assumption. In the second one, the assumption of a single active source per time-frequency bin is replaced by a multi-source assumption, as employed in [21, 22] for source separation. Third, we evaluate the impact of an angular spectrum-based initialization for clustering-based algorithms by comparing it with random initialization. Finally, we provide the largest-scale evaluation of multi-source TDOA estimation approaches to date by comparing the proposed approaches and five state-of-the-art approaches on 1482 different configurations.

The rest of the paper is organized as follows. In section 2, a short review of existing angular spectrum-based methods is given and the proposed SNR-based approaches are presented. Existing and proposed clustering-based methods are presented in section 3. The experimental evaluation is detailed in section 4, and conclusions are drawn in section 5.

## 2 Angular spectrum-based methods

The principle of angular spectrum-based methods consists in constructing a function  $\phi(\tau)$  of TDOA  $\tau$  whose peaks indicate the TDOAs of the sources. This is commonly achieved as follows. A local angular spectrum (or coherence measure)  $\phi(t, f, \tau)$  is computed in each time-frequency bin  $(t, f)$  for all discrete values of  $\tau$  lying on a uniform grid in the range of possible TDOAs. This function is chosen so that it is likely to exhibit large values for the TDOAs of the sources which are active in this time-frequency bin. In order to robustify the estimation process and to overcome the spatial aliasing ambiguity occurring at high frequencies, the function  $\phi(t, f, \tau)$  is summed over all frequencies. Then, it is reduced to a single dimension to obtain the angular spectrum from which the TDOAs are estimated. This is typically done by summing it over all time frames [3]:

$$\phi^{\text{sum}}(\tau) = \sum_{t=1}^T \sum_{f=1}^F \phi(t, f, \tau). \quad (3)$$

A limitation of this approach is that it makes it difficult to localize a source that is active only within few time frames due to the integration of irrelevant information when the source is inactive. This can be addressed by taking the

maximum over all time frames instead [13]:

$$\phi^{\max}(\tau) = \max_t \sum_{f=1}^F \phi(t, f, \tau). \quad (4)$$

Existing methods differ by the definition of the local angular spectrum function  $\phi(t, f, \tau)$  and the choice of the pooling function, i.e., “sum” (3) or “max” (4).

## 2.1 Popular existing methods

Existing methods typically extract the spatial information in time-frequency bin  $(t, f)$  from the *empirical covariance matrix*  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$  of the input signal, which can be computed in the neighborhood of each time-frequency bin  $(t, f)$  as [23]

$$\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) = \frac{\sum_{t', f'} w(t' - t, f' - f) \mathbf{x}(t', f') \mathbf{x}(t', f')^H}{\sum_{t', f'} w(t' - t, f' - f)}, \quad (5)$$

where  $w$  is a time-frequency windowing function defining the size and the shape of the neighborhood, and  $(\cdot)^H$  denotes the Hermitian transposition operator.

The generalized cross-correlation with phase transform (GCC-PHAT) method [3] is certainly the most popular angular spectrum-based method. Based on the assumption that the direct sound of one source predominates in each time-frequency bin, the TDOA of this source  $\tau$  is estimated from the phase difference between the two channels represented by the argument of  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)_{1,2}$ <sup>1</sup>. Indeed, this phase difference is expected to be close to  $2\pi f\tau$  modulo  $2\pi$ . The local angular spectrum is then defined as

$$\phi^{\text{GCC}}(t, f, \tau) = \Re \left( \frac{\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)_{1,2}}{|\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)_{1,2}|} e^{-2i\pi f\tau} \right), \quad (6)$$

where  $\Re(z)$  denotes the real part of a complex number  $z$ . In [24], it is proposed to use a non-linear function  $\rho$  of  $\phi^{\text{GCC}}(t, f, \tau)$  defined by  $\rho(u) = 1 - \tanh(\alpha\sqrt{1-u})$  in order to emphasize the large values of  $\phi^{\text{GCC}}(t, f, \tau)$ .

In multiple signal classification (MUSIC) [5], under the same assumption that one source is predominant in each time-frequency bin  $(t, f)$ , the local angular spectrum is computed as a measure of fit between the steering vector  $\mathbf{d}(f, \tau)$  and the first principal component  $\mathbf{v}(t, f)$  of  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$ :

$$\phi^{\text{MUSIC}}(t, f, \tau) = \left( 1 - \frac{1}{\sqrt{2}} \mathbf{d}(f, \tau)^H \mathbf{v}(t, f) \right)^{-1}, \quad (7)$$

where  $\mathbf{d}(f, \tau)$  is defined by (2).

Nesta *et al.* [12] propose a method relaxing the assumption of one predominant source in each time-frequency bin. The time-frequency plane is split into time-frequency blocks, and it is assumed that there are at most two predominant sources in each block. Then, Independent Component Analysis (ICA) is

<sup>1</sup>Here  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)_{i,j}$  denotes the  $(i, j)$ -th element of matrix  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$ .

applied in each time-frequency block  $(t, f)$  to obtain two amplitude-normalized mixing coefficients  $r_1(t, f)$  and  $r_2(t, f)$  (see [12] for details) that are likely to be close to  $e^{-2i\pi f\tau_1}$  and  $e^{-2i\pi f\tau_2}$  up to a permutation, where  $\tau_1$  and  $\tau_2$  are the TDOAs of the two predominant sources in the considered block. The local angular spectrum of this method called *cumulative state coherence transform* (cSCT) is given by

$$\phi^{\text{cSCT}}(t, f, \tau) = \sum_{j=1}^2 \rho \left( \frac{1}{2} |e^{-2i\pi f\tau} - r_j(t, f)| \right), \quad (8)$$

where  $\rho(u) = 1 - \tanh(\alpha\sqrt{u})$ .

## 2.2 Proposed methods

All existing approaches, except the cSCT, use the assumption that in each time-frequency bin, only one source is predominant. For cSCT, this assumption is replaced by that of at most two predominant sources. Both assumptions do not hold exactly for most audio data [23, 25]. Thus, some of the estimated local angular spectra  $\phi(t, f, \tau)$  do not represent any “true” TDOA possibly leading to poor estimation of the TDOAs. A bounded time-frequency weighting function based on interchannel correlation has been proposed in [26] to give more weight to the time-frequency bins involving a single predominant source. However, this function leads to overestimate the weight at low frequencies where interchannel correlation is large regardless of the number of active sources. Inspired by [17], where it was done for instantaneous mixtures, we propose to use the SNR as an unbounded measure to determine whether the information contained in a time-frequency bin results from a single source. We propose three methods to estimate the SNR below.

### 2.2.1 SNR estimation by beamforming

In each time-frequency bin, we define the SNR in the direction corresponding to the TDOA  $\tau$  by the ratio between the sound power in this direction and the residual power. We estimate the power in the direction corresponding to the TDOA  $\tau$  by the *Minimum Variance Distortionless Response* (MVDR) beamformer [27]:

$$P(t, f, \tau) = \left( \mathbf{d}(f, \tau)^H \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)^{-1} \mathbf{d}(f, \tau) \right)^{-1}, \quad (9)$$

where  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$  and  $\mathbf{d}(f, \tau)$  are computed by (5) and (2), respectively. Then, we compute the residual power by subtracting the estimated power in the direction from an estimate of the total power:  $\frac{1}{2} \text{tr} \left( \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) \right) - P(t, f, \tau)$ . Finally, we define the SNR in this direction as

$$\phi^{\text{MVDR}}(t, f, \tau) = \frac{\left( \mathbf{d}(f, \tau)^H \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)^{-1} \mathbf{d}(f, \tau) \right)^{-1}}{\frac{1}{2} \text{tr} \left( \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) \right) - \left( \mathbf{d}(f, \tau)^H \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)^{-1} \mathbf{d}(f, \tau) \right)^{-1}}. \quad (10)$$

Our preliminary experiments showed that this function overestimates the SNR at low frequencies, where phase differences are small, regardless of the number of active sources. As a consequence, the values of  $\phi^{\text{MVDR}}(t, f, \tau)$  at low frequencies mask the values at higher frequencies (see Fig. 1, left).

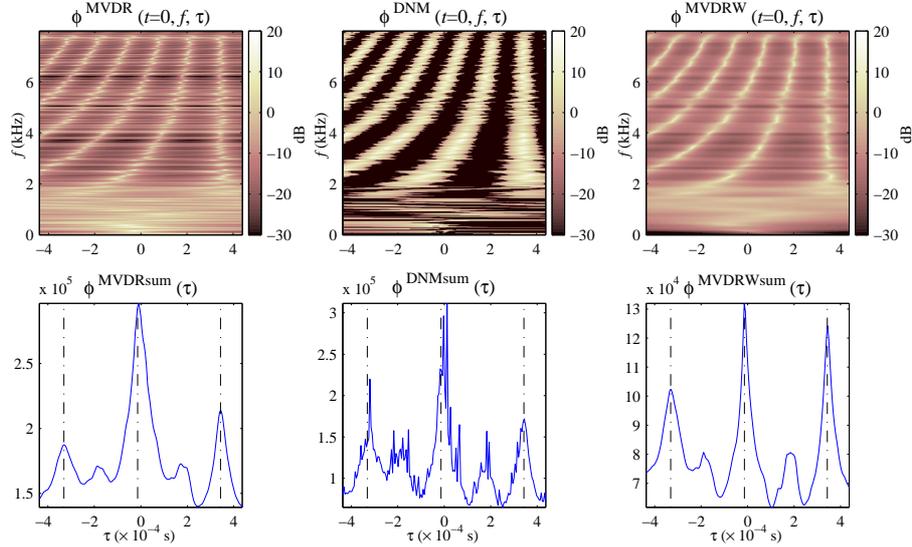


Figure 1: Local angular spectra  $\phi(t, f, \tau)$  plotted at time  $t = 0$  (top) and pooled angular spectra  $\phi^{\text{sum}}(\tau)$  (bottom) for  $\phi^{\text{MVDR}}$  (left),  $\phi^{\text{DNM}}$  (center) and  $\phi^{\text{MVDRW}}$  (right), computed from the mixture of three female speech sources placed at 50 cm from the center of the microphone pair, with  $d = 15$  cm and a reverberation time of 500 ms. Dash-dotted lines indicate the true TDOAs.

### 2.2.2 SNR estimation under a diffuse noise model

To address this problem, we propose to jointly estimate the sound power in direction  $\tau$  and the power of the residual signal by using a diffuse noise model. We assume that in each time-frequency bin only one source  $s(t, f)$  of TDOA  $\tau$  is predominant, i.e., the mixing model (1) becomes

$$\mathbf{x}(t, f) = \mathbf{d}(f, \tau)s(t, f) + \mathbf{b}(t, f). \quad (11)$$

Moreover, we assume that  $s(t, f)$  and  $\mathbf{b}(t, f)$  follow independent zero-mean Gaussian distributions with, respectively, variance  $v^s(t, f, \tau)$  and covariance  $v^b(t, f, \tau)\mathbf{\Psi}(f)$ . The variances  $v^s(t, f, \tau), v^b(t, f, \tau) > 0$  represent the source and noise power, and  $\mathbf{\Psi}(f)$  is the covariance matrix of a gain-normalized diffuse noise [6, 20]:

$$\mathbf{\Psi}(f) = \begin{pmatrix} 1 & \text{sinc}(2\pi f \frac{d}{c}) \\ \text{sinc}(2\pi f \frac{d}{c}) & 1 \end{pmatrix}, \quad (12)$$

with  $d$  being the distance between the two microphones (in meters),  $c$  the speed of sound (in meters/sec.), and  $\text{sinc}(y) = \frac{\sin(y)}{y}$  the cardinal sine function. With these assumptions, it can be shown from (11) that the covariance matrix of the mixture in the time-frequency bin  $(t, f)$  and for TDOA  $\tau$  can be expressed as

$$\mathbf{R}_{\mathbf{xx}}(t, f, \tau) = v^s(t, f, \tau)\mathbf{d}(f, \tau)\mathbf{d}^H(f, \tau) + v^b(t, f, \tau)\mathbf{\Psi}(f), \quad (13)$$

and the log-likelihood of  $\mathbf{x}(t, f)$  can be written as:

$$\begin{aligned} \log p(\mathbf{x}(t, f)) &= \log N(\mathbf{x}(t, f); 0, \mathbf{R}_{\mathbf{xx}}(t, f, \tau)) \triangleq \\ &= -\text{tr} \left( \mathbf{R}_{\mathbf{xx}}^{-1}(t, f, \tau) \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) \right) - \log \det (\pi \mathbf{R}_{\mathbf{xx}}(t, f, \tau)). \end{aligned} \quad (14)$$

Using a closed form solution from [20], we estimate  $v^s(t, f, \tau)$  and  $v^b(t, f, \tau)$  in the maximum likelihood sense as

$$\begin{pmatrix} v^s(t, f, \tau) \\ v^b(t, f, \tau) \end{pmatrix} = (\text{diag}(\mathbf{\Lambda}_1), \text{diag}(\mathbf{\Lambda}_2))^{-1} \text{diag}(\mathbf{A}^{-1} \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) (\mathbf{A}^H)^{-1}), \quad (15)$$

where  $(\mathbf{Y}, \mathbf{Z})$  denotes the concatenation of matrices (or vectors)  $\mathbf{Y}$  and  $\mathbf{Z}$ ,  $\text{diag}(\mathbf{Y})$  denotes the column vector of diagonal entries of matrix  $\mathbf{Y}$ ,  $\mathbf{A}$  is the matrix whose columns are the eigenvectors of  $\mathbf{d}(f, \tau) \mathbf{d}^H(f, \tau) \mathbf{\Psi}^{-1}(f)$ , and  $\mathbf{\Lambda}_1$ ,  $\mathbf{\Lambda}_2$  are equal respectively to  $\mathbf{A}^{-1} \mathbf{d}(f, \tau) \mathbf{d}^H(f, \tau) (\mathbf{A}^H)^{-1}$  and  $\mathbf{A}^{-1} \mathbf{\Psi}(f) (\mathbf{A}^H)^{-1}$ . Non-negativity of  $v^s(t, f, \tau)$  and  $v^b(t, f, \tau)$  is imposed by setting  $v^s(t, f, \tau)$  to zero and  $v^b(t, f, \tau)$  to  $\frac{1}{2} \text{tr} \left( \mathbf{\Psi}^{-1}(f) \widehat{\mathbf{R}}_{\mathbf{xx}}(t, f) \right)$  when  $v^s(t, f, \tau)$  or  $v^b(t, f, \tau)$  resulting from (15) is negative [20]. Finally, the SNR in the time-frequency bin  $(t, f)$  and for TDOA  $\tau$  is computed as

$$\phi^{\text{DNM}}(t, f, \tau) = \frac{v^s(t, f, \tau)}{v^b(t, f, \tau)}. \quad (16)$$

A drawback of this method is that the resulting angular spectra are often not smooth enough to obtain a good estimate of the TDOAs (see Fig. 1, center).

### 2.2.3 SNR estimation by frequency weighted beamforming

To take advantage of both the smoothness of  $\phi^{\text{MVDR}}$  (see Fig. 1, left) and the global shape of  $\phi^{\text{DNM}}$  (see Fig. 1, center), we combine these two methods by expressing a relationship between them, assuming that the input signal consists of a single source of TDOA  $\tau = 0$  and diffuse noise. In other words, we replace in (10) the empirical covariance matrix  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$  by the covariance matrix  $\mathbf{R}_{\mathbf{xx}}(t, f, \tau)$  in (13), and we get

$$\phi^{\text{MVDR}}(t, f, \tau) = \frac{1 + 2v^s(t, f, \tau)/v^b(t, f, \tau)}{1 - \text{sinc}(2\pi f \frac{d}{c})}. \quad (17)$$

By inverting (17), we estimate the SNR as:

$$\phi^{\text{MVDRW}}(t, f, \tau) = \frac{v^s(t, f, \tau)}{v^b(t, f, \tau)} = w_d(f) \phi^{\text{MVDR}}(t, f, \tau) - \frac{1}{2}, \quad (18)$$

where  $w_d(f) = \frac{1}{2}(\text{sinc}(2\pi f \frac{d}{c}) - 1)$  is a weighting factor depending on the frequency and the microphone spacing  $d$ . This factor reduces the impact of low frequencies, as it can be seen from Figure 2 and from Figure 1 (top right) below 1 KHz. It can be also noted from Figure 1 that  $\phi^{\text{MVDRW}}$  is as smooth as  $\phi^{\text{MVDR}}$ , and, at the same time, it has sharper peaks.

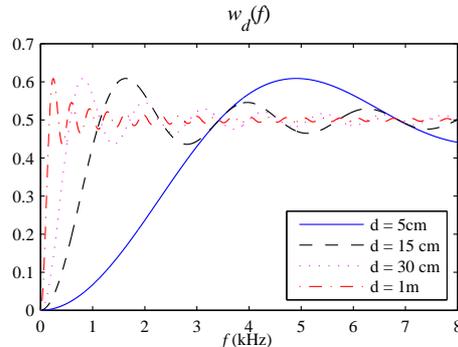


Figure 2: Weighting factor  $w_d(f)$  for microphone spacings  $d = 5$  cm, 15 cm, 30 cm and 1 m.

### 3 Clustering-based methods

In angular spectrum-based methods a measure of source activity (e.g., the SNR) can be exploited, but the estimation of this measure does not rely on any estimates of the TDOAs. In contrast, the principle of clustering-based methods consists of estimating this measure given some estimates of the TDOAs, then reestimating the TDOAs relying on this measure, and so on. This results in iterating between the following two steps:

- re-estimate the contribution of each source to each time-frequency bin (*clusters*), given current estimates of the TDOAs,
- re-estimate the TDOAs, given current estimates of the clusters.

#### 3.1 Popular existing methods

Sawada *et al.* [10] propose to perform the clustering in a hard manner, by associating each time-frequency bin to the closest source  $n_{tf}$  in the sense of the Euclidean distance between its steering vector and the phase and amplitude-normalized observation. Each TDOA  $\tau_n$  is then reestimated from the time-frequency bins of the corresponding cluster only.

It is also possible to perform the clustering in a soft manner using a probabilistic setting. In [6], the clustering is performed with an Expectation-Maximization (EM) algorithm, assuming that the mixture in each time-frequency bin  $(t, f)$  consists of only one predominant source  $n_{tf}$  and a diffuse noise:

$$\mathbf{x}(t, f) = s_{n_{tf}}(t, f)\mathbf{d}(f, \tau_{n_{tf}}) + \mathbf{b}(t, f). \quad (19)$$

The source coefficients  $s_{n_{tf}}(t, f)$  are considered as deterministic parameters and the noise term  $\mathbf{b}(t, f)$  is modeled as a zero-mean Gaussian random vector of covariance matrix  $v^b\Psi(f)$ , where  $\Psi(f)$  is given by (12) and  $v^b$  is a constant parameter.

## 3.2 Proposed methods

### 3.2.1 EM algorithm with one predominant source in each time-frequency bin

In line with [6], we assume that in each time-frequency bin  $(t, f)$ , there is only one predominant source  $n_{tf}$  and a diffuse noise  $\mathbf{b}(t, f)$ . However, in contrast to [6], we assume that both  $s_{n_{tf}}(t, f)$  and  $\mathbf{b}(t, f)$  follow independent zero-mean Gaussian distributions with covariances  $v_{n_{tf}}^s(t, f)$  and  $v_{n_{tf}}^b(t, f)\Psi(f)$ , where  $v_{n_{tf}}^s(t, f)$  and  $v_{n_{tf}}^b(t, f)$  represent respectively the source and the noise variances, and  $\Psi(f)$  is defined by (12). Given the predominant source index  $n_{tf}$ , the mixture covariance matrix is  $\mathbf{R}_{\mathbf{x}\mathbf{x}, n_{tf}}(t, f) = \Xi_{\mathbf{x}}(f, \tau_{n_{tf}}, v_{n_{tf}}^s(t, f), v_{n_{tf}}^b(t, f))$ , where

$$\Xi_{\mathbf{x}}(f, \tau, v^s, v^b) \triangleq v^s \mathbf{d}(f, \tau) \mathbf{d}^H(f, \tau) + v^b \Psi(f). \quad (20)$$

The set of parameters to be estimated is

$$\theta = \left\{ \{\tau_n\}_{n=1}^N, \{v_n^s(t, f), v_n^b(t, f)\}_{n,t,f=1}^{N,T,F} \right\}. \quad (21)$$

Under the above assumptions and assuming a uniform prior over the source indices, the observation  $\mathbf{x}(t, f)$  follows the Gaussian mixture model (GMM) distribution

$$p(\mathbf{x}(t, f) | \theta) = \sum_n \frac{1}{N} p(\mathbf{x}(t, f) | \tau_n, v_n^s(t, f), v_n^b(t, f)), \quad (22)$$

where

$$p(\mathbf{x}(t, f) | \tau, v^s, v^b) = N(\mathbf{x}(t, f); 0, \Xi_{\mathbf{x}}(f, \tau, v^s, v^b)), \quad (23)$$

and  $N(\mathbf{x}(t, f); 0, \Xi_{\mathbf{x}}(f, \tau, v^s, v^b))$  is defined by (14).

We use an EM algorithm [28] to estimate the parameters  $\theta$  in the maximum likelihood sense, considering the set of predominant source indices  $\{n_{tf}\}_{t,f}$  as latent data. The updates of the resulting algorithm and some hints for its derivation are given in Appendix A.

### 3.2.2 EM algorithm with multiple sources in each time-frequency bin

As it was already mentioned in section 2.2, the assumption of only one predominant source does not hold exactly for most audio data. Several works [23, 25, 21, 22] have shown that relaxing this assumption can be very beneficial for audio source separation. Thus, we here investigate whether such an approach could be beneficial for multi-source localization. The model presented below is mostly inspired by the models proposed in [21, 22]. We consider that all sources can be present in each time-frequency bin and model the mixing process as

$$\mathbf{x}(t, f) = \mathbf{D}(f, \boldsymbol{\tau}) \mathbf{s}(t, f) + \mathbf{b}(t, f), \quad (24)$$

where  $\boldsymbol{\tau} = [\tau_1, \dots, \tau_N]$  is the vector of TDOAs,  $\mathbf{D}(f, \boldsymbol{\tau}) = [\mathbf{d}(f, \tau_1), \dots, \mathbf{d}(f, \tau_N)]$ ,  $\mathbf{s}(t, f) = [s_1(t, f), \dots, s_N(t, f)]^T$  and  $\mathbf{b}(t, f) = [b_1(t, f), b_2(t, f)]^T$ .

We assume that  $\mathbf{b}(t, f)$  and  $\mathbf{s}(t, f)$  follow zero-mean Gaussian distributions with covariance matrices respectively equal to  $v^b(t, f)\Psi(f)$  and  $\mathbf{R}_{\mathbf{ss}}(t, f)$ , where  $\mathbf{R}_{\mathbf{ss}}(t, f)$  is the diagonal matrix with  $n$ -th diagonal entry being equal to  $v_n^s(t, f)$ .

With these assumptions the observation  $\mathbf{x}(t, f)$  follows a zero-mean Gaussian distribution with covariance matrix

$$\mathbf{R}_{\mathbf{xx}}(t, f) = \mathbf{D}(f, \boldsymbol{\tau})\mathbf{R}_{\mathbf{ss}}(t, f)\mathbf{D}^H(f, \boldsymbol{\tau}) + v^b(t, f)\boldsymbol{\Psi}(f). \quad (25)$$

The set of parameters to be estimated is

$$\theta = \left\{ \{\tau_n\}_{n=1}^N, \{v_n^s(t, f)\}_{n,t,f=1}^{N,T,F}, \{v^b(t, f)\}_{t,f}^{T,F} \right\}. \quad (26)$$

To estimate the parameters in the maximum likelihood sense we derive an EM algorithm, considering the sources  $\{\mathbf{s}(t, f)\}_{t,f}$  as latent data. The algorithm is summarized in Appendix B. We consider four variants of the algorithm:

1. “EM-multi (TF noise)”: the noise variances  $v^b(t, f)$  are unconstrained,
2. “EM-multi (F noise)”: in line with [21], the noise variances  $v^b(t, f)$  are constrained to be constant over time, i.e.,  $v^b(t, f) = v^b(f)$ ,
3. “EM-multi (const noise)”: in line with [6], the noise variances  $v^b(t, f)$  are constrained to be constant over time and frequency, i.e.,  $v^b(t, f) = v^b$ ,
4. “EM-multi (no noise)”: in line with [10], the noise variances  $v^b(t, f)$  are fixed to a small positive value  $\varepsilon^{\text{fix}}$ .

## 4 Evaluation

### 4.1 Data

The experimental evaluation was carried out on a large number of mixtures of male speech, female speech and music sources taken from the database of the 2008 Signal Separation Evaluation Campaign (SiSEC) [29] “under-determined speech and music mixtures” task. Mixing filters were simulated with the *Roomsimove* Toolbox<sup>2</sup> for a rectangular room of dimensions 4.45 m  $\times$  3.55 m  $\times$  2.5 m and omnidirectional microphones. We considered all possible combinations of the following parameters:

- Number of sources  $N$ : from 2 to 6.
- Reverberation time  $\text{RT}_{60}$ <sup>3</sup>: 50 ms, 100 ms, 150 ms, 250 ms, 500 ms, 750 ms.
- Microphone spacing  $d$ : 5 cm, 15 cm, 30 cm, 1 m.
- Distance between the sources and the center of the microphone pair: 20 cm, 50 cm, 1 m, 2 m.

<sup>2</sup>E. Vincent, D. Campbell, Roomsimove, a Matlab toolbox for the computation of simulated room impulse responses for moving sources, <http://www.irisa.fr/metiss/members/evincen/software>.

<sup>3</sup> $\text{RT}_{60}$  is the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

- Angular position of the sources: between 3 and 5 randomly generated scenarios according to the number of sources, i.e., 5 scenarios for 2 sources, 4 for 3 - 4 sources, and 3 for 5 - 6 sources. Moreover, the scenarios were generated with the following restrictions: the *Directions Of Arrival* (DOAs) cannot be smaller (greater) than 30 degrees (150 degrees) and the absolute difference between DOAs of any pair of sources cannot be smaller than 15 degrees for 2-5 sources and 10 degrees for 6 sources.
- Three source types (female speech, male speech and music).

We only kept situations for which the distance between the sources and the microphone pair is lower than  $0.8d$  to ensure that the far-field assumption holds, so that the relation between a TDOA  $\tau$  and its corresponding DOA  $\eta$  (in degrees) can be expressed by

$$\tau = \frac{d}{c} \cos\left(2\pi \frac{\eta}{360}\right). \quad (27)$$

This resulted in a total of 4446 mixtures. All the signals were sampled at 16 kHz.

## 4.2 Evaluation measures

### 4.2.1 Localization

Evaluation in terms of localization consists in measuring the capacity of the methods to estimate the true TDOAs with some tolerance. Since we do not attempt to estimate the number of sources  $N$ , we apply the considered methods for all possible numbers of sources  $J$  from 1 to 20. For a given  $J$ , we select the  $J$  highest peaks of the angular spectra and run the clustering algorithms with  $J$  clusters. We then evaluate the list of  $J$  estimated TDOAs compared to the list of  $N$  true TDOAs in terms of recall, precision and F-measure. An estimated TDOA  $\hat{\tau}$  is considered as correctly estimated if it is close enough to the true TDOA  $\tau$ , in the sense that the corresponding DOAs  $\hat{\eta}$  and  $\eta$ , computed by inverting (27), differ by less than 5 degrees modulo 180 degrees. Denoting by  $I_J$  the number of correctly estimated TDOAs, recall, precision and F-measure are defined by [30]

$$\text{Recall}(J) = \frac{I_J}{N}, \quad (28)$$

$$\text{Precision}(J) = \frac{I_J}{J}, \quad (29)$$

$$\text{F-measure}(J) = 2 \times \frac{\text{Recall}(J) \times \text{Precision}(J)}{\text{Recall}(J) + \text{Precision}(J)}. \quad (30)$$

Figure 3 shows the average recall, precision and F-measure of several angular spectrum-based methods over all mixtures with  $N = 3$  sources. The results are represented as a function of the assumed number of sources  $J$ . As expected, the recall increases and the precision decreases with  $J$ , while the F-measure is maximum for some  $J = J_{\text{opt}}$ .  $J_{\text{opt}}$  equals to the number of sources  $N$  for some algorithms, but not for all. Thus,  $J_{\text{opt}}$  depends on the algorithm  $\mathcal{A}$  and on the number of sources  $N$ . We also noticed that it depends on the microphone spacing  $d$  but that it is insensitive to the other parameters. Thus, for

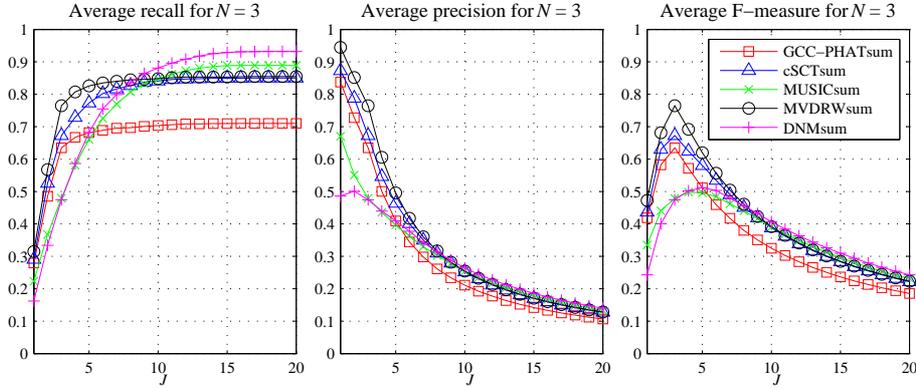


Figure 3: Average recall (left), precision (center) and F-measure (right) as functions of  $J$  for several angular spectrum-based methods and for all mixtures with  $N = 3$  sources.

each algorithm  $\mathcal{A}$ , each number of sources  $N$  and each microphone spacing  $d$ , we compute  $J_{\text{opt}}(\mathcal{A}, N, d)$  that maximizes the average F-measure over the corresponding mixtures. In the rest of this section, we assume that  $J$  is fixed to  $J_{\text{opt}}(\mathcal{A}, N, d)$  for all mixtures.

#### 4.2.2 Accuracy

When the TDOAs are correctly estimated, one would like to know how accurate these estimates are. Let the TDOAs of all the mixtures considered be enumerated through as  $\{\tau^i\}_{i=1}^I$ . Let  $\mathcal{I} \subset \{1, \dots, I\}$  denote the subset of TDOAs that are correctly estimated by all the methods under comparison. We define the average accuracy as follows:

$$\text{Accuracy} = \frac{1}{\sum_{i \in \mathcal{I}} N_i^{-1}} \sum_{i \in \mathcal{I}} N_i^{-1} |(\eta^i - \hat{\eta}^i) \bmod 180|, \quad (31)$$

where  $\eta^i$  and  $\hat{\eta}^i$  are the DOAs computed from the true  $\tau^i$  and estimated  $\hat{\tau}^i$  TDOAs using (27), and  $N_i$  denotes the number of sources in the corresponding mixture.

#### 4.3 Parameters

All the methods evaluated below had exactly the same front-end to compute the empirical covariance matrices  $\hat{\mathbf{R}}_{\mathbf{x}\mathbf{x}}(t, f)$  (5), except the cSCT [12], for which we used an implementation provided by the author of [12]. The STFT was computed with half-overlapping sine windows of length 1024. The time-frequency windowing function  $w$  in (5) was a  $3 \times 3$  Hanning window.

#### 4.4 Angular spectrum-based methods

We consider four state-of-the-art local angular spectra: GCC-PHAT [3], a version of GCC-PHAT with a non-linear function [24] (denoted as GCC-NONLIN),

Local angular spectrum function	Recall (0 to 1)		Precision (0 to 1)		F-measure (0 to 1)		Accuracy (degrees)	
	$\phi^{\text{sum}}$	$\phi^{\text{max}}$	$\phi^{\text{sum}}$	$\phi^{\text{max}}$	$\phi^{\text{sum}}$	$\phi^{\text{max}}$	$\phi^{\text{sum}}$	$\phi^{\text{max}}$
GCC-PHAT [3]	0.61	0.84	0.72	<b>0.85</b>	0.65	0.84	0.93	<b>0.37</b>
GCC-NONLIN [24]	0.67	<b>0.85</b>	0.73	<b>0.85</b>	0.69	<b>0.85</b>	0.66	0.39
MUSIC [5]	0.65	0.61	0.41	0.30	0.50	0.40	-	-
cSCT [12]	0.65	-	0.68	-	0.66	-	0.79	-
MVDR	0.63	0.64	0.66	0.41	0.64	0.50	-	-
DNM	0.70	0.65	0.43	0.33	0.53	0.43	-	-
MVDRW	<b>0.73</b>	0.80	<b>0.75</b>	0.80	<b>0.74</b>	0.80	<b>0.51</b>	0.46

Table 1: Angular spectrum-based methods: average results for all 4446 mixtures. The average accuracy was computed over the 7167 TDOAs correctly estimated by the methods with average F-measure greater than 0.64 (from a total of 16614).

MUSIC [5] and cSCT [12], and the three proposed SNR-based local angular spectra: MVDR, DNM and MVDRW (see Sec. 2.2). All seven local angular spectra were evaluated with both  $\phi^{\text{sum}}(\cdot)$  (3) and  $\phi^{\text{max}}(\cdot)$  (4) pooling functions, except the cSCT [12]<sup>4</sup>. Note also that for all the methods, in line with [24], we do not select peaks that differ by less than 5 degrees from a higher peak.

The results in terms of average recall, precision, F-measure and accuracy are reported in Table 1. Since the average accuracy is computed over the TDOAs that are correctly estimated by all the compared methods, as explained in section 4.2.2, we have selected for this comparison the methods that performed better than 0.64 in terms of average F-measure. This was done in order to avoid a comparison over a very small subset of TDOAs.

Analyzing the results in terms of F-measure, we see that for the “sum” pooling function MVDRW outperforms all other methods, which confirms our findings in [20]. Using the “max” pooling function instead of “sum” improves MVDRW. However, it improves even more GCC-PHAT and GCC-NONLIN, which perform best in the end. The results in terms of average accuracy appear correlated with the average F-measure, and GCC-PHATmax is the most accurate on average.

We then investigate the behaviour of the average F-measure as a function of the microphone spacing, the reverberation time and the number of sources. For this comparison we chose five methods among the best ones, namely GCC-NONLINsum, GCC-NONLINmax, cSCTsum, MVDRWsum and MVDRWmax. The results are shown in Figure 4. While the reverberation time and the number of sources have little influence on the ranking of different methods, the microphone spacing has. Note that in case of the “sum” pooling function the advantage of MVDRW against GCC-NONLIN was essentially due to its better performance for the 5 cm microphone spacing. Using the “max” pooling function improves GCC-NONLIN for all microphone spacings, but it improves MVDRW only for small spacings, while leading to a big performance degradation for large spacings. However, our proposed MVDRWmax method outperforms GCC-NONLINmax for the 5 cm microphone spacing.

<sup>4</sup>Since we have only an implementation of the cSCT [12] given us by the author of [12] without the corresponding sources, we were not able injecting the “max” pooling function into this implementation.

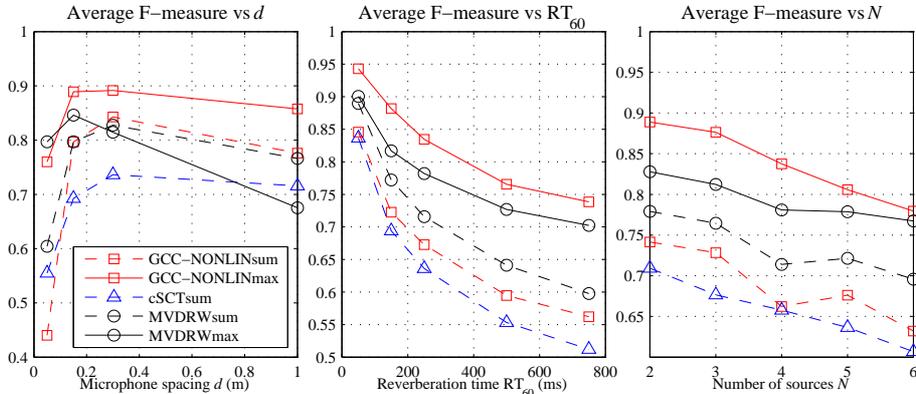


Figure 4: Angular spectrum-based methods: average F-measure as a function of the microphone spacing  $d$  (left), the reverberation time  $RT_{60}$  (center), and the number of sources  $N$  (right).

#### 4.5 Clustering-based methods

The evaluation of clustering-based methods, as compared to that of angular spectrum-based methods, is much more computationally expensive within our evaluation framework. Indeed, as explained in Sec. 4.2.1, each clustering-based method must be run 20 times (for  $J$  from 1 to 20) with several EM iterations. Since we have not noticed any influence of signal type on the ranking of angular spectrum-based methods, we retain only the female speech mixtures for this evaluation.

We consider two state-of-the-art methods by Sawada *et al.* [10] and Izumi *et al.* [6], and the four of the proposed methods: the “EM-predom” method introduced in Sec. 3.2.1 and the variants of the “EM-multi” method introduced in Sec. 3.2.2, except the “EM-multi (TF noise)” method that was discarded because of its poor performance in some preliminary tests. To evaluate the impact of initialization on the performance, each method was run twice: (i) with randomly initialized TDOAs and (ii) with initial TDOAs estimated by GCC-NONLINmax, i.e., the best angular spectrum-based method. For each mixture, the random initial TDOAs were drawn from the uniform distribution on the interval  $[-d/c, d/c]$ . For each method the source and the noise variances  $v_n^s(t, f)$  and  $v_n^b(t, f)$  were initialized by some constant values  $v_n^{s,init}$  and  $v_n^{b,init}$  adjusted during preliminary tests. Each method was run for 10 iterations.

The results in terms of average evaluation measures, including those of the initialization (noted by “None”), are summarized in Table 2. Almost all methods slightly improve the average F-measure, as compared to the random initialization. However, when run with the initialization provided by GCC-NONLINmax all the clustering-based methods lead to some performance degradation. The smallest degradation is obtained by the two noise-free methods, i.e., Sawada *et al.* [10] and “EM-multi (no noise)”, possibly because they are the most constrained and do not change the initial TDOAs so much. The average accuracy appears correlated with the average F-measure again, and GCC-NONLINmax is the most accurate on average.

Clustering algorithm	Recall (0 to 1)		Precision (0 to 1)		F-measure (0 to 1)		Accuracy (degrees)	
	rand	init	rand	init	rand	init	rand	init
None	0.53	<b>0.90</b>	0.24	<b>0.90</b>	0.32	<b>0.90</b>	-	<b>0.45</b>
Sawada <i>et al.</i> [10]	0.65	0.88	<b>0.76</b>	0.87	0.49	0.87	-	0.63
Izumi <i>et al.</i> [6]	0.29	0.30	0.48	0.51	0.30	0.35	-	-
EM-predom	0.68	0.73	0.54	0.56	<b>0.50</b>	0.62	-	-
EM-multi (F noise)	0.73	0.69	0.45	0.69	0.37	0.68	-	1.15
EM-multi (const noise)	0.78	0.75	0.49	0.70	0.39	0.72	-	1.19
EM-multi (no noise)	<b>0.92</b>	0.88	0.47	0.88	0.33	0.88	-	0.47

Table 2: Clustering-based methods: average results for 1482 female speech mixtures. Notations: “rand” means initializing clustering by random TDOAs, “init” means initializing it by TDOAs estimated with GCC-NONLINmax, and “None” means evaluating the initialization without running any clustering algorithm. The average accuracy was computed over the 2708 TDOAs correctly estimated by the methods with average F-measure greater than 0.64 (from a total of 5538).

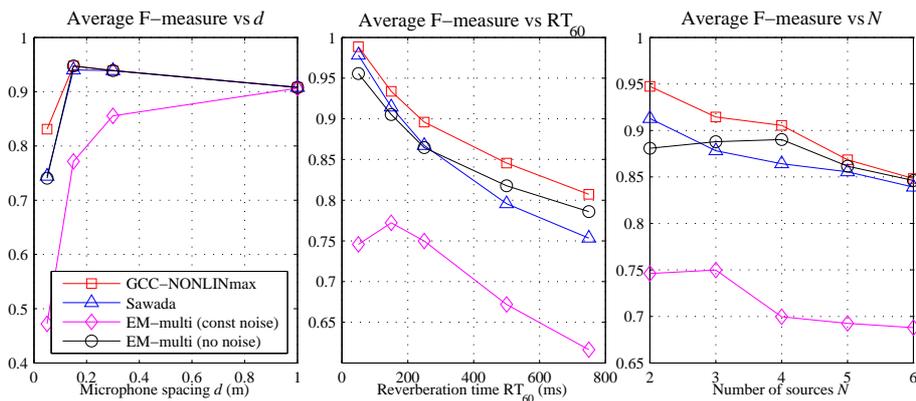


Figure 5: Clustering-based methods: average F-measure as a function of microphone spacing  $d$  (left), reverberation time  $RT_{60}$  (center), and number of sources  $N$  (right).

We choose the four best methods, namely GCC-NONLINmax, Sawada *et al.* [10], “EM-multi (const noise)” and “EM-multi (no noise)”, to investigate the behaviour of the average F-measure as a function of different parameters. The results are shown in Figure 5. We see that the clustering-based methods, as compared to GCC-NONLINmax used for the initialization, decrease the average F-measure for all considered conditions. Moreover, the drop in performance is mostly pronounced for small microphone spacings, while for high and moderate spacings (15 cm, 30 cm and 1 m) there is almost no performance decrease for the two noise-free clustering-based methods.

## 5 Conclusion

In this paper we introduced several multi-source TDOA estimation methods based on angular spectra and clustering. The common motivation behind all these methods is to go beyond the assumption of a single predominant source per time-frequency bin. This is achieved by either using the SNR as an unbounded measure of source activity or by exploiting directly the multi-source hypothesis within a probabilistic model. We evaluated the proposed and the five most popular state-of-the-art approaches on 1482 different configurations. To our knowledge, this is by far the largest-scale evaluation of multi-source TDOA estimation approaches to date.

Among all angular spectrum-based methods, evaluated with both the “sum” and “max” pooling functions, a version of the state-of-the-art GCC-PHAT method with the “max” pooling function gave the best results on average and the best results for large and moderate microphone spacings (15 cm, 30cm and 1 m). The proposed SNR-based MVDRW method with the “max” pooling function outperformed GCC-PHAT for 5 cm microphone spacing.

All clustering-based methods were evaluated with both random TDOA initialization and an initialization provided by the best angular spectrum-based method. First, we showed that it is very important to initialize the clustering algorithms with “good” TDOA estimates, since random initialization leads to really poor performance. Second, we observed that none of the clustering-based methods was able to improve the TDOA estimates compared to the best angular spectrum-based method.

As for further research, the role of the “max” pooling function, its success and its potential limits should be better understood, so as to be able to propose more powerful non-linear pooling functions. Moreover, we still believe in the potential of clustering-based methods. One way to improve them would be to replace the models of the direct and the reverberant parts we proposed by more structured audio-specific models, e.g., as in [21]. An alternative way would be to try injecting the “max” pooling function within the probabilistic model behind clustering.

## A Updates for the EM algorithm with one pre-dominant source in each time-frequency bin

An EM algorithm [28] is an iterative algorithm consisting in updating the parameters  $\theta^{(l)}$  at every iteration  $l$  as follows:

$$\theta^{(l)} = \arg \max_{\theta} Q(\theta, \theta^{(l-1)}), \quad (32)$$

where  $Q(\theta, \theta')$  is the so-called auxiliary function. Under the assumptions of Section 3.2.1 this function is equal (up to some additive constant) to

$$Q(\theta, \theta') \stackrel{c}{=} \sum_{n,t,f} \gamma_n(t, f) p(\mathbf{x}(t, f) | \tau_n, v_n^s(t, f), v_n^b(t, f)), \quad (33)$$

with

$$\gamma_n(t, f) \triangleq p(n_{tf} = n | \mathbf{x}(t, f), \theta') \quad (34)$$

$$\propto p(\mathbf{x}(t, f) | \tau'_n, v_n^{s'}(t, f), v_n^{b'}(t, f)), \quad (35)$$

where, for any parameter  $z$  from  $\theta$ ,  $z'$  denotes the corresponding parameter from  $\theta'$  and  $p(\mathbf{x}(t, f)|\tau, v^s, v^b)$  is defined by (23).

As before, the TDOAs are only estimated on the discrete uniform grid in the range of possible TDOAs. Let us denote this grid by  $\Gamma$ . One iteration of the EM parameter updates optimizing (32) for the auxiliary function (33) consists of the following steps:

1. Estimate  $\hat{v}_n^s(t, f, \tau)$  and  $\hat{v}_n^b(t, f, \tau)$  in the maximum likelihood sense for all time-frequency bins and all possible TDOAs  $\tau \in \Gamma$  using (15)<sup>5</sup>.
2. Compute the posterior cluster probabilities  $\gamma_n(t, f)$  as in (35) and normalize them so that  $\sum_n \gamma_n(t, f) = 1$ .
3. Update the TDOAs  $\tau_n$  as follows:

$$\tau_n = \arg \max_{\tau \in \Gamma} \sum_{t, f} \gamma_n(t, f) \log p(\mathbf{x}(t, f)|\tau, \hat{v}_n^s(t, f, \tau), \hat{v}_n^b(t, f, \tau)), \quad (36)$$

where  $p(\mathbf{x}(t, f)|\tau, \hat{v}_n^s(t, f, \tau), \hat{v}_n^b(t, f, \tau))$  is defined by (23).

4. Update the source and noise variances by setting  $v_n^s(t, f) = \hat{v}_n^s(t, f, \tau_n)$  and  $v_n^b(t, f) = \hat{v}_n^b(t, f, \tau_n)$ .
5. Set  $\theta' = \theta$ .

## B Updates for the EM algorithm with multiple sources in each time-frequency bin

Performing some derivations analogous to those from [21] and [22], it can be shown that the auxiliary function  $Q(\theta, \theta')$  for the EM algorithm under assumptions of Section 3.2.2 is equal (up to some additive constant) to:

$$Q(\theta, \theta') \stackrel{c}{=} - \sum_{n, t, f} \left( \log v_n^s(t, f) + \frac{\hat{\mathbf{R}}_{\text{ss}}(t, f)_{n, n}}{v_n^s(t, f)} \right) - \sum_{t, f} \left( 2 \log v^b(t, f) + \frac{\mathcal{M}(t, f, \boldsymbol{\tau}, \theta')}{v^b(t, f)} \right), \quad (37)$$

where

$$\begin{aligned} \mathcal{M}(t, f, \boldsymbol{\tau}, \theta') \triangleq & \text{tr} \left[ \boldsymbol{\Psi}^{-1}(f) \hat{\mathbf{R}}_{\text{xx}}(t, f) - \boldsymbol{\Psi}^{-1}(f) \mathbf{D}(f, \boldsymbol{\tau}) \hat{\mathbf{R}}_{\text{xs}}^H(t, f) \right. \\ & \left. - \boldsymbol{\Psi}^{-1}(f) \hat{\mathbf{R}}_{\text{xs}}(t, f) \mathbf{D}^H(f, \boldsymbol{\tau}) + \boldsymbol{\Psi}^{-1}(f) \mathbf{D}(f, \boldsymbol{\tau}) \hat{\mathbf{R}}_{\text{ss}}(t, f) \mathbf{D}^H(f, \boldsymbol{\tau}) \right], \end{aligned} \quad (38)$$

$$\hat{\mathbf{R}}_{\text{xs}}(t, f) = \hat{\mathbf{R}}_{\text{xx}}(t, f) \mathbf{G}_s^H(t, f), \quad (39)$$

$$\hat{\mathbf{R}}_{\text{ss}}(t, f) = \mathbf{G}_s(t, f) \hat{\mathbf{R}}_{\text{xs}}(t, f) + [\mathbf{I}_N - \mathbf{G}_s(t, f) \mathbf{D}(f, \boldsymbol{\tau}')] \mathbf{R}'_{\text{ss}}(t, f), \quad (40)$$

<sup>5</sup>Note that in practice this needs to be done only once for all iterations.

with  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$  defined by (5), and

$$\mathbf{G}_{\mathbf{s}}(t, f) = \mathbf{R}'_{\mathbf{ss}}(t, f) \mathbf{D}^H(f, \boldsymbol{\tau}') (\mathbf{R}'_{\mathbf{xx}}(t, f))^{-1}, \quad (41)$$

$$\mathbf{R}'_{\mathbf{xx}}(t, f) = \mathbf{D}(f, \boldsymbol{\tau}') \mathbf{R}'_{\mathbf{ss}}(t, f) \mathbf{D}^H(f, \boldsymbol{\tau}') + v^{b'}(t, f) \boldsymbol{\Psi}(f). \quad (42)$$

One iteration of the EM parameter updates optimizing (32) for the auxiliary function (37) consists of the following steps:

1. Compute  $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$ ,  $\widehat{\mathbf{R}}_{\mathbf{xs}}(t, f)$  and  $\widehat{\mathbf{R}}_{\mathbf{ss}}(t, f)$  using (5), (39) and (40).
2. Update the source variances as  $v_n^s(t, f) = \widehat{\mathbf{R}}_{\mathbf{ss}}(t, f)_{n,n}$ .
3. Update the TDOAs as  $\tau_n = \tau_n^*$ , where:

$$\tau_n^* = \arg \min_{\tau_n \in \Gamma} \sum_{t, f} \frac{1}{v^b(t, f)} \mathcal{M}(t, f, \boldsymbol{\tau}, \theta'), \quad (43)$$

and  $\mathcal{M}(t, f, \boldsymbol{\tau}, \theta')$  is defined by (38). More precisely, each TDOA  $\tau_n$  ( $n = 1, \dots, N$ ) is updated in turn, while keeping the other TDOAs  $\{\tau_m\}_{m \neq n}$  fixed<sup>6</sup>.

4. Update noise variances:

$$v^b(t, f) = \frac{1}{2 \cdot \#\mathcal{J}(t, f)} \sum_{(\tilde{t}, \tilde{f}) \in \mathcal{J}(t, f)} \mathcal{M}(\tilde{t}, \tilde{f}, \boldsymbol{\tau}, \theta'), \quad (44)$$

with  $\mathcal{J}(t, f) \subset \{1, \dots, T\} \times \{1, \dots, F\}$  denoting the subset of time-frequency bins where  $v^b(t, f)$  is considered constant and  $\#\mathcal{J}(t, f)$  denoting the number of elements in this subset. This formulation allows the implementation of the first three variants of the algorithm mentioned at the end of section 3.2.2. To implement the fourth variant one just needs to skip the noise variance update. s

5. Set  $\theta' = \theta$ .

## Acknowledgment

The authors would like to thank Francesco Nesta for sharing his cSCT implementation, and Nobutaka Ito for discussion during the course of this work.

<sup>6</sup>Both the usage of the noise variances  $v^b(t, f)$  in the updates of TDOAs (43) and the alternating nature of these updates do not guarantee the maximization of the auxiliary function, as in (32). However, they guarantee its non-decrease, i.e.,  $Q(\theta^{(l)}, \theta^{(l-1)}) \geq Q(\theta^{(l-1)}, \theta^{(l-1)})$ . Thus, the resulting algorithm is rather a Generalized EM (GEM) algorithm [28]. Updating TDOAs jointly, instead of alternatively, is possible as well, but it is avoided here, since it would lead to a computational complexity growing exponentially with the number of sources.

## References

- [1] W. Kellermann, *Acoustic Echo Cancellation for Beamforming Microphone Arrays*, Springer, Berlin, 2001, pp. 281–306.
- [2] S. Makino, H. Sawada, T.-W. Lee, *Blind Speech Separation*, Springer, 2007.
- [3] C. Knapp, G. Carter, The generalized cross-correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech and Signal Processing* 24 (4) (1976) 320–327.
- [4] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, *IEEE Transactions on Signal Processing* 52 (7) (2004) 1830–1847.
- [5] R. Schmidt, Multiple emitter location and signal parameter estimation, *IEEE Transactions on Antennas and Propagation* 34 (3) (1986) 276–280.
- [6] Y. Izumi, N. Ono, S. Sagayama, Sparseness-based 2ch BSS using the EM algorithm in reverberant environment, in: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 147–150.
- [7] C. Faller, J. Merimaa, Source localization in complex listening situations: Selection of binaural cues based on interaural coherence, *Journal of the Acoustical Society of America* 116 (5) (2004) 3075–3089.
- [8] H. Viste, G. Evangelista, On the use of spatial cues to improve binaural source separation, in: *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2003, pp. 209–213.
- [9] J. Mouba, S. Marchand, A source localization/separation/respatialization system based on unsupervised classification of interaural cues, in: *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2006, pp. 233–238.
- [10] H. Sawada, S. Araki, R. Mukai, S. Makino, Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation, *IEEE Transactions on Audio, Speech, and Language Processing* 15 (5) (2007) 1592–1604.
- [11] M. Mandel, D. Ellis, T. Jebara, An EM algorithm for localizing multiple sound sources in reverberant environments, in: *Neural Information Processing Systems*, 2006, pp. 953–960.
- [12] F. Nesta, P. Svaizer, M. Omologo, Cumulative state coherence transform for a robust two-channel multiple source localization, in: *Proc. 8th Int Conf on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 290–297.
- [13] Z. El Chami, A. Guerin, A. Pham, C. Servière, A phase-based dual microphone method to count and locate audio sources in reverberant rooms, in: *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 209–212.

- 
- [14] C. Liu, B. C. Wheeler, Jr, R. C. Bilger, C. R. Lansing, A. S. Feng, Localization of multiple sound sources with two microphones, *Journal of the Acoustical Society of America* 108 (4) (2000) 1888–1905.
- [15] A. Brutti, M. Omologo, P. Svaizer, Comparison between different sound source localization techniques based on a real data collection, in: *Proc. 2nd Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2008, pp. 69–72.
- [16] B. Loesch, B. Yang, Comparison of different algorithms for acoustic source localization, in: *ITG Fachtagung Sprachkommunikation*, 2010.
- [17] S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture, *IEEE Transactions on Signal Processing* 58 (1) (2010) 121–133.
- [18] Y. Deville, M. Puigt, Temporal and time-frequency correlation-based blind source separation methods. Part I: Determined and underdetermined linear instantaneous mixtures, *Signal Processing* 87 (2007) 374–407.
- [19] P. Bofill, Identifying single source data for mixing matrix estimation in instantaneous blind source separation, in: *18th Int. Conf. on Artificial Neural Networks (ICANN)*, 2008, pp. 759–767.
- [20] C. Blandin, E. Vincent, A. Ozerov, Multi-source TDOA estimation using SNR-based angular spectra, in: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, to appear.
- [21] A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, *IEEE Transactions on Audio, Speech, and Language Processing* 18 (3) (2010) 550–563.
- [22] N. Q. K. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model, *IEEE Transactions on Audio, Speech, and Language Processing* 18 (7) (2010) 1830–1840.
- [23] E. Vincent, S. Arberet, R. Gribonval, Underdetermined instantaneous audio source separation via local Gaussian modeling, in: *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2009, pp. 775 – 782.
- [24] B. Loesch, B. Yang, Blind source separation based on time-frequency sparseness in the presence of spatial aliasing., in: *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 1–8.
- [25] S. Arberet, A. Ozerov, R. Gribonval, F. Bimbot, Blind spectral-GMM estimation for underdetermined instantaneous audio source separation, in: *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2009, pp. 751–758.
- [26] N. Roman, D. Wang, G. Brown, Speech segregation based on sound localization, *Journal of the Acoustical Society of America* 114 (4) (2003) 2236–2252.

- [27] H. Krim, M. Viberg, Two decades of array signal processing research: the parametric approach, *IEEE Signal Processing Magazine* 13 (4) (1996) 67–94.
- [28] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* 39, No. 1 (1977) 1–38.
- [29] E. Vincent, S. Araki, P. Bofilld, The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation, in: *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.
- [30] C. van Rijsbergen, *Information retrieval*, 2nd Edition, Butterworths, London, UK, 1979.



---

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

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

---

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