# SOUND SOURCE SEPARATION IN THE HIGHER ORDER AMBISONICS DOMAIN

*Mohammed Hafsati*
b<>com, Rennes, France

*Nicolas Epain*
b<>com, Rennes, France

*Rémi Gribonval*
Univ Rennes, Inria, CNRS, IRISA, France

*Nancy Bertin*
Univ Rennes, Inria, CNRS, IRISA, France

**ABSTRACT**

In this article we investigate how the *local Gaussian model* (LGM) can be applied to separate sound sources in the higher-order ambisonics (HOA) domain. First, we show that in the HOA domain, the mathematical formalism of the local Gaussian model remains the same as in the microphone domain. Second, using an off-the shelf source separation toolbox (FASST) based on the local Gaussian model, we validate the efficiency of the approach in the HOA domain by comparing the performance of toolbox in the HOA domain with its performance in the microphone domain. To do this we discuss and run some simulations to ensure a fair comparison. Third, we check the efficiency of the local Gaussian model compared to other available source separation techniques in the HOA domain. Simulation results show that separating sources in the HOA domain results in a 1 to 12 dB increase in signal-to-distortion ratio, compared to the microphone domain.

Multichannel source separation, local Gaussian model, Wiener filtering, 3D audio, Higher Order Ambisonics (HOA).

## 1. INTRODUCTION

There is an increasing interest in new, immersive forms of media, such as 360-degree videos and Virtual Reality (VR) experiences. Producing media experiences of this kind requires new techniques and workflows on both the video and audio sides. In particular, the feeling of immersion that is sought after by VR spectators highly depends on the quality of the audio rendering. In order for the experience to be convincing, sounds must be binauralized according to the spectator's location and orientation relative to the different sound sources.

Among the various 3D-audio technologies available, Higher-Order Ambisonics (HOA) [1, 2, 3, 4] has become the *de facto* standard for 360-degree video soundtracks. This is primarily because HOA provides a panoramic representation of the sound field, which can easily be rotated in accordance with the listener's head orientation prior to being played back. Another significant advantage of the HOA representation is that it is straightforward to record HOA sound scenes using relatively compact Spherical Microphone Arrays (SMAs). In contrast to 360-degree videos, however, VR experiences not only allow the spectator to look in any direction, but also to navigate through the virtual environment. This means that the spectator's perspective of the scene may change over time in a manner that cannot be modeled as a simple rotation effect. For instance, the spectator can move toward one of the sound sources, which should translate in this source being louder compared to other sources.

One possible approach to simulate a movement through the scene is to interpolate between several HOA representations corresponding to different points of view [5, 6]. However, in practice this requires to use several SMAs, which may be impossible. Another possibility consists in decomposing one HOA scene into directional components, which typically correspond to sound sources, and changing their directions and gains according to the listener's movements [7, 8, 9]. In addition to being more practical than the interpolation method, this approach was shown to yield a better listening experience [8]. The quality of the navigation effect obtained with the decomposition approach ultimately depends on the accuracy of the sound source separation. In previous work [9], we showed that sound sources could be separated using a simple beamforming technique, which relies solely on spatial information. In the presence of complex sound scenes, however, the quality of the separation could be improved using multi-channel source separation algorithms, such as those based on the so-called *local Gaussian model* [10, 11, 12].

In this article, we investigate the ability of the local Gaussian model to handle the informed source separation problem (knowing directions of arrival) directly in the HOA domain. In Section. 2, we recall the model in the microphone domain and derive its equivalent in the HOA domain. This allows us to perform experiments with an off-the-shelf source separation toolbox, the Flexible Audio Source Separation Toolbox (FASST) [13], presented in Section 3. Experimental results are presented and discussed in Section 5; First, on a small dataset, we investigate the performance of FASST with respect to of the number of channels in the HOA domain and the number of microphones in the microphone domain, and select a number of channels/microphones yielding a fair comparison between the HOA domain and microphone domain. Second, on a large scale dataset, we measure the performance of FASST in both domains, and compare them to each other. Third, we compare the performance of FASST with different types of beamformers. We conclude in Section 6 on the validity of the local Gaussian model and its competitive performance when applied to HOA signals, including in challenging situations such as highly reverberating environments or close source positions.

## 2. MIXTURE MODELS

### 2.1. Microphone domain

The source separation problem consists in estimating the contribution $\mathbf{c}_{j,t} \in \mathbb{R}^I$ of each source $j = 1, ..., J$ in each microphone $i = 1, ..., I$ and at each time instant $t = 1, ..., T$. In the absence of noise, the mixture can be written as:

$$\mathbf{x}_t = \sum_{j=1}^{J} \mathbf{c}_{j,t}\,, \tag{1}$$

where $\mathbf{x}_t = [x_{1,t}, \ldots, x_{I,t}]^\mathsf{T} \in \mathbb{R}^I$ are microphone array signals. In a reverberant environment, under the hypothesis of point sources, the source signal $s_{j,t}$ can be related to its contribution $\mathbf{c}_{j,t}$ through:

$$\mathbf{c}_{j,t} = [\alpha_{ij} * s_j]_t\,, \tag{2}$$

where $*$ denotes the convolution product, $\alpha_{ij}$ is the impulse response of the mixing filter between the source $j$, and the microphone $i$. Now, under the narrow-band approximation, and assuming the mixing filters are time invariant, the Short-Time Fourier Transform (STFT) of the microphone signals is given by:

$$\mathbf{x}_{f,n} = \sum_{j=1}^{J} \mathbf{A}_{j,f}\mathbf{s}_{j,fn}\,, \tag{3}$$

where $f$ and $n$ denote the frequency bin and time-frame index, respectively. $\mathbf{A}_f \in \mathbb{C}^{I\times J}$ contains the frequency responses $a_{ij,f}$ of the filters $a_{ij}(t)$, and embeds information on the source directions of arrival (DOA).

The source separation problem as defined in Eq.(1) can be addressed using the multichannel Wiener filtering framework, which will be presented with more details in Sec. 3. This framework requires the selection of a distribution model for the variables to estimate. For simplicity we use the local Gaussian model presented in [14]:

$$\forall f \in [1, F]\,, n \in [1, N]\,, \quad \mathbf{c}_{j,f,n} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}})\,, \tag{4}$$

where $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = \mathbb{E}\left[\mathbf{c}_{j,f,n}\,\mathbf{c}_{j,f,n}^\mathsf{H}\right]$ is the covariance matrix of the contribution of the $j$-th source to every microphone at frequency $f$ and time frame $n$. In line with the literature, this matrix can be further decomposed as the product of a scalar spectral part, $v_{j,f,n}$, with a time-invariant spatial matrix, $\mathbf{R}_{\mathbf{c}_{j,f}}$, as follows: $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = v_{j,f,n}\,\mathbf{R}_{\mathbf{c}_{j,f}}$. Notably, the so-called spatial covariance matrix $\mathbf{R}_{\mathbf{c}_{j,f}}$ respects the relation $\mathbf{R}_{\mathbf{c}_{j,f}} = \mathbf{A}_{j,f}\,\mathbf{A}_{j,f}^\mathsf{H}$ when the assumptions of Eq. (3) hold.

### 2.2. HOA domain

In the Higher-Order Ambisonic (HOA) framework, the sound field is decomposed over a basis of spherical harmonic functions. The HOA signals, $\mathbf{z}_t$, are typically obtained by applying a set of finite impulse response filters, known as encoding filters, to the signals recorded by a spherical microphone array [15]. Thus, assuming the encoding filters are short enough, the vector of the HOA signal STFTs $\mathbf{z}_{f,n} \in \mathbb{C}^M$ is given by:

$$\mathbf{z}_{f,n} = \mathbf{E}_f\,\mathbf{x}_{f,n}, \tag{5}$$

where $\mathbf{E}_f$ is the matrix of the encoding filter frequency responses. Using Eq. (1), we can now model the HOA mixture as follows:

$$\mathbf{z}_{f,n} = \sum_{j=1}^{J} \mathbf{E}_f\,\mathbf{c}_{j,f,n}, \tag{6}$$

and identify the contribution of the $j$-th source to the different HOA channels as:

$$\mathbf{b}_{j,f,n} = \mathbf{E}_f\,\mathbf{c}_{j,f,n}. \tag{7}$$

As is the case in the microphone domain, in the ambisonic domain source separation consists in estimating the contribution of every source to every channel $\mathbf{b}_{j,f,n}$, which can be solved using a Wiener filtering approach. To this aim we assume the following local Gaussian model:

$$\mathbf{b}_{j,f,n} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{b}_{j,f,n}}). \tag{8}$$

Similar to the microphone domain, the covariance $\mathbf{\Sigma}_{\mathbf{b}_{j,f,n}}$ can be further decomposed into a spectral part, $v_{j,f,n}$, and a spatial covariance matrix given by:

$$\mathbf{R}_{\mathbf{b}_{j,f}} = \mathbf{E}_f\,\mathbf{R}_{\mathbf{c}_{j,f}}\,\mathbf{E}_f^\mathsf{H}\,. \tag{9}$$

## 3. SOURCE SEPARATION WITH FASST

The multi-channel source separation problem can be solved by looking for the filter that minimizes the expected squared error for every source $j$ and every time frequency bin $(f, n)$:

$$\forall j \in [1, J], f \in [1, F] \text{ and } n \in [1, N],$$
$$\mathbf{W}_{j,f,n} = \underset{\mathbf{W}}{\operatorname{argmin}}\, \mathbb{E}\left[\|\mathbf{c}_{j,f,n} - \mathbf{W}_{j,f,n}\,\mathbf{x}_{f,n}\|_2^2\right]\,. \tag{10}$$

The filter $\mathbf{W}_{j,f,n}$ is known as the multichannel Wiener filter and is given by:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{(\mathbf{c}_{j,f,n},\,\mathbf{x}_{f,n})}\,\mathbf{\Sigma}^{-1}_{(\mathbf{x}_{f,n},\,\mathbf{x}_{f,n})}\,, \tag{11}$$

where the matrices $\mathbf{\Sigma}_{(\mathbf{x}_{f,n},\,\mathbf{x}_{f,n})}$ and $\mathbf{\Sigma}_{(\mathbf{c}_{j,f,n},\,\mathbf{x}_{f,n})}$, represent the covariance of the mixture $\mathbf{x}_{f,n}$ and the cross-correlation between the vectors $\mathbf{c}_{j,f,n}$ and $\mathbf{x}_{f,n}$, respectively.

From Eq. (4), and assuming the sources are statistically independent, the Wiener filter can be simplified as:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}\left(\sum_{j'=1}^{J} \mathbf{\Sigma}_{\mathbf{c}_{j',f,n}}\right)^{-1}\,. \tag{12}$$

Thus, the source separation problem reduces to the problem of estimating the covariance matrices $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}$ or, equivalently in the HOA domain, $\mathbf{\Sigma}_{\mathbf{b}_{j,f,n}}$. Each source contribution is obtained by applying element-wise its corresponding Wiener filter to the mixture: $\mathbf{c}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{x}_{fn}$, respectively, $\mathbf{b}_{j,f,n} = \mathbf{W}_{j,f,n}\mathbf{z}_{j,f,n}$ in the HOA domain. and finally using overlap-add to reconstruct the time-domain signal.

In this work we use the flexible audio source separation toolbox (FASST) [13, 16], a software toolbox which allows the estimation of these parameters and apply the subsequent Wiener filter. In FASST the parameters are estimated by maximizing the log-likelihood of the observations with an Expectation-Maximization (EM) algorithm, and a multi-channel non negative matrix factorization (NMF) model can be enforced on the source covariances $\Sigma_{c_{j,f,n}}$ [11].

## 4. EXPERIMENTAL PROTOCOL

### 4.1. Dataset

In order to evaluate the source separation performance, we built a dataset as follows. First, fifty songs were picked randomly from the Mixing Secret Dataset (MSD100)[1]. In the MSD100 database, each song consists of four sound sources (voice, bass, drums and "others") provided as separate tracks.

In this work, microphone array recordings were then simulated using MCRoomSim [17], a room acoustics simulation software. This software calculates impulse responses modeling the acoustic propagation between acoustic sources and sensors in reverberant environments. A total of 16 simulations were run, corresponding to four rooms and four source configurations. The eight rooms had the same dimensions, 10 m $\times$ 8 m $\times$ 3 m, but different wall absorption coefficients, which resulted in the following reverberation times: 0 s, 0.2 s, 0.4 s and 0.7 s. The four source configurations are illustrated in Fig. 1. In every simulation the microphone array was modeled to match the characteristics of the Eigenmike[2] and was located at the same position in the room. In order to calculate the microphone mixtures, for each song and each of the 32 conditions the separate source tracks were then convolved with the simulated impulse responses and summed with each other.

We then built two different inputs for source separation: a microphone mixture $\mathbf{x}$ obtained using every sensor of the Eigenmike (32 channels), and a fourth order ambisonic mixture $\mathbf{b}$ (25 channels) obtained by applying encoding filters to the microphone mixture $\mathbf{x}$.

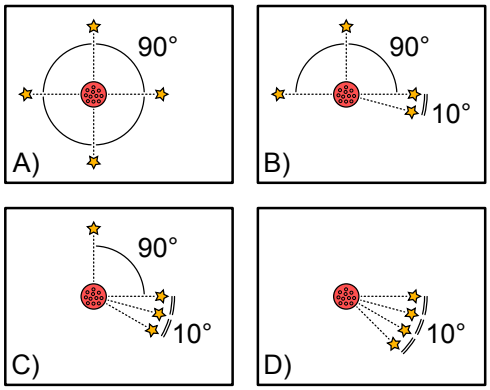


**Fig. 1**: The four sound source configurations considered in our simulations. Note: stars represent sound source locations. All the sources and the spherical microphone array are at the same height. Across all the examples, the position of sources with similar type was the same.

### 4.2. Evaluation criteria

In order to validate the adaptability of FASST in the HOA domain, we propose to compare its performance to that obtained by applying FASST in the microphone domain. A fair comparison requires to compute the chosen performance measures in the same domain. However, some information is lost when converting microphone signals to HOA signals, therefore it is impossible to convert the separated HOA signals back to the microphone domain for comparison. To alleviate this issue, instead of computing the evaluation measures in terms of the contribution of each source in each channel/microphone (FASST's outputs), we propose to compute them in terms of sound objects.

We obtain sound objects by applying beamforming to the signals separated by FASST. For simplicity, we use the beamforming technique known as the matched filter for both the microphone and HOA domains. For the j-th sound source, the sound object is calculated by projecting the separated signals onto the steering vector corresponding to a plane wave incoming from the source direction,$(\theta_j, \phi_j)$. In other words, the estimated source object $j$ in the microphone and HOA domains are calculated as follows:

$$\hat{s}^{\text{Mic}}_{j,f,n} = \frac{\mathbf{a}^H_{j,f}}{||\mathbf{a}_{j,f}||^2}\hat{\mathbf{c}}_{j,f,n} \quad (13)$$

$$\hat{s}^{\text{HOA}}_{j,f,n} = \frac{\mathbf{y}^H_{j,f}}{||\mathbf{y}_{j,f}||^2}\hat{\mathbf{b}}_{j,f,n} \quad (14)$$

where $\mathbf{y}_j$ corresponds to the spherical harmonic vector evaluated at the direction of arrival of the source $j$. We then compare the estimated sound object signals to reference signals, which we define as the sound objects obtained by applying the same beamforming to the actual (i.e. oracle) sound source contributions. In other words for the source $j$ the reference signal in the microphone and HOA domains are given by:

$$s^{\text{Mic}}_{j,f,n} = \frac{\mathbf{a}^H_{j,f}}{||\mathbf{a}_{j,f}||^2}\mathbf{c}_{j,f,n} \quad (15)$$

$$s^{\text{HOA}}_{j,f,n} = \frac{\mathbf{y}^H_{j,f}}{||\mathbf{y}_{j,f}||^2}\mathbf{b}_{j,f,n} \quad (16)$$

Lastly, we assess the source separation performance by comparing the signals given by Eq. (13), and Eq. (14) to the ones given by Eq. (15) and Eq. (16), respectively, using the following performance measures [18]: Signal to Distortion Ratio (SDR), Signal to Artifact Ratio (SAR), and Signal to Interference Ratio (SIR). These measures are then calculated with the BSS-eval toolbox[3] [19].

### 4.3. Evaluated methods

The first method we examine consists in applying FASST to the HOA mixtures (see Sec. 3.) We compare it to the equivalent microphone-domain mixtures with the same number of channels. We also compare the performance obtained by using FASST in the HOA domain with that of two different beamformers.

The first beamformer has already been introduced in Section. 4.2. It is the matched filter beamformer(PWD), but this time applied directly to the HOA mixture, which is given by:

$$\bar{s}^{\text{HOA}}_{j,f,n} = \frac{\mathbf{y}^H_{j,f}}{||\mathbf{y}_{j,f}||^2}\mathbf{z}_{f,n}. \quad (17)$$

The second beam former, which we refer as the pseudo-inverse beamfomer, consists in multiplying the HOA signals with the pseudo-inverse of the matrix containing the steering vectors for the directions of the sources. The resulting beamformer is a particular

[1] https://sisec.inria.fr/sisec-2015/2015-professionally-produced-music-recordings/

[2] https://mhacoustics.com/products

[3] BSS-eval version 3.0 for Matlab, http://bass-db.gforge.inria.fr/bss_eval/

case of the Linearly-Constrained Minimum-Variance beamformer (LCMV). It is given by:

$$\hat{\bar{s}}^{\mathrm{HOA}}_{j,f,n} = (\mathbf{Y}^H\mathbf{Y})^{-1}\mathbf{Y}^H\mathbf{z}_{f,n}, \tag{18}$$

where the matrix $\mathbf{Y}$ contains the spherical harmonic vectors of the sources direction of arrivals.

### 4.4. FASST parametrization and initialization

The FASST toolbox requires choosing configuration parameters, as well as providing initial values for the covariance matrices $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}$ in Eq. (12). Tab. 1 summarizes the parameters used for all experiments. Further, in order to match the scene configuration, the number of sources was fixed to 4 in the anechoic condition and 5 in reverberant conditions, where we observed that it was beneficial to add a source accounting for diffuse noise or late reverberation.

| Transform type | STFT |
|---|---|
| Sampling frequency | 44100 Hz |
| Window length | 69 ms (3072 samples) |
| NMF rank | 16 |
| Stopping criterion | 150 iterations |

**Table 1**: FASST parameters

In FASST, covariances are decomposed into a spectral part and a spatial part, and the spectral part further modeled by NMF, as proposed by [16]. For each of the first 4 sources, the spatial covariance was initialized to the rank-1 matrices $\mathbf{R}_f^{\mathrm{HOA}} = \mathbf{y}_j\mathbf{y}_j^{\mathsf{H}}$ and $\mathbf{R}_f^{\mathrm{Mic}} = \mathbf{a}_{j,f}\mathbf{a}_{j,f}^{\mathsf{H}}$ for the HOA and microphone domain, respectively. In the microphone domain, the steering vectors $\mathbf{a}_{j,f}$ were derived from the microphone array characteristics and source locations. In the HOA domain, the steering vectors $\mathbf{y}_j$ were derived as the vector of the first nine spherical harmonic functions evaluated in the source directions. In the reverberant case, the fifth source was assumed to have a full-rank spatial covariance, which was was initialized to the identity matrix. Lastly, regarding the spectral part of the covariance, NMF factors were initialized as random numbers.

## 5. VALIDATION OF THE APPROACH

As explained before the main goal is to validate experimentally the local Gaussian model assumption for source separation in the HOA domain. To this aim the performances obtained using FASST in the HOA domain are compared to that obtained in the microphone domain. However, we first need to select a number of microphone channels and HOA signals that ensures a fair comparison between the two methods. Thus, we first investigate the influence of the number of microphone channels and HOA signals on the source separation performance for a fraction of the dataset. The results of this study, presented in Sec. 5.1, indicate that it is fair to compare the results obtained with 9 HOA signals (order 2 HOA signals) with that obtained using 9 microphone channels. In Sec. 5.2 we present further experimental results obtained using the selected number of channels for the entire dataset.

### 5.1. Selection of the number of microphones/channels

The computational cost of FASST depends primarily on the square of the number of channels of the mixture, and considering the size of our dataset, it is important to spare time and resources in the main experiment that will soon be described. A naive approach would be to adopt a lower HOA order $L < 4$, and consider on the one hand HOA mixtures with $M = (L+1)^2$ channels, and on the other one, the same mixtures given by a sub-antenna of the Eigenmike, where the number of the chosen capsules is $I = (L+1)^2$.

However, one could argue that while HOA mixtures are obtained by considering all of the capsules of the Eigenmike, the microphone mixtures are given by only $M = (L+1)^2$ selected microphones, and therefore, the comparison could be considered unfair. To clarify this point, we begin our experiments by measuring the source separation performance in both domains when varying respectively the number of channels and the number of microphones. This preliminary experiment is done on a small proportion of the created dataset (see below).

First, in the microphone domain we considered different sub antennas from the Eigenmike where the capsules were selected in order to be distributed regularly on the sphere. The considered numbers of microphones are $I = 4, 9, 12, 16, 25, 32$ (the numbers $4, 9, 16, 25$ were chosen to match the number of possible channels in the HOA domain, the number $12$ is considered because the chosen capsules can be regularly distributed in the best way to cover the sphere). Second, HOA signals make sense if they are grouped by order $L$, each order $L$ corresponding to a number of channels $M = (L+1)^2$. We have already at our disposal the 4th order signals (25 channels) by encoding the information provided by the 32 capsules of the Eigenmike. In order to have the first, the second, and the third order we have to simply truncate respectively the 25 HOA signals to the first $M = 4, 9, 16$ channels. Considering the selected capsules in the microphone domain and the truncation of the signals in the HOA domain, from our data set we considered randomly 160 mixtures, all the listed time reverberations and sound source configurations were considered.

We applied FASST to the different mixtures, considering the sources DOA known, the initialization and the parametrization of the toolbox are given in Sec. 4.4. The results in terms of SDR, SIR and SAR are given in Fig. 2.

In the microphone domain, we observe that the SIR tends to improve by 0.07 dB in average when increasing the number of microphones, the SAR tends to decrease, when it comes to the SDR we observe that it increases slightly by 0.02 dB in average until 9 microphones and drops after. In the HOA domain, we observe an improvement of all performance measures when increasing the number of channels. We can clearly see that adding more microphones doesn't improve the source separation performance in the microphone domain. As a conclusion it is unnecessary to add more microphones in the microphone domain, and therefore the comparison of FASST's performance between the HOA domain and the microphone domain is fair if the number of channels/microphones is equal to $I = M = 9$.

### 5.2. Extensive experiments with 9 microphones/channels

In the following the considered number of channels is equal to the considered number of microphones $I = M = 9$. In the microphone domain the selected capsules are given in Table. 2. More

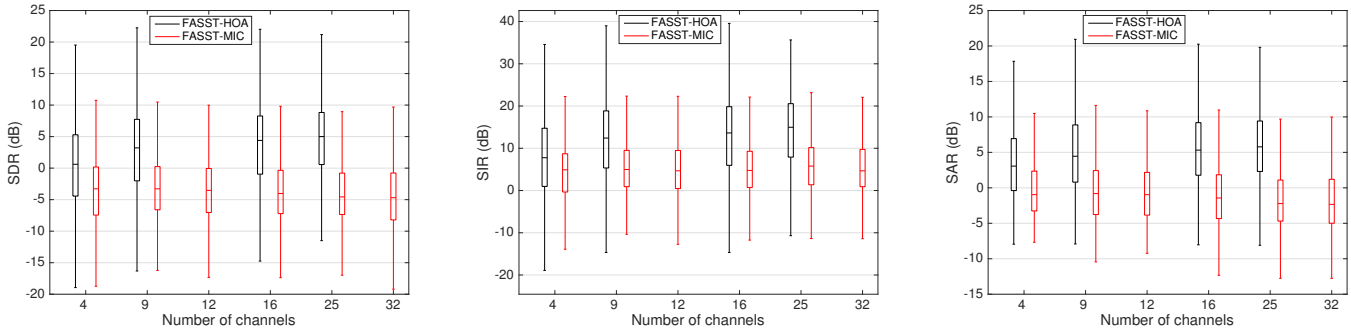


**Fig. 2**: Comparing FASST performance in regards of number the used microphones/channels.

information about the angular position of the Eigenmike's capsules can be found in [20].

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | 0 | 35 | -58 | -31 | 0 | -58 | 35 | 69 | -32 |
| $\phi$ | -32 | 45 | 0 | 90 | 212 | 180 | 135 | 269 | -90 |

**Table 2**: Elevation ($\theta$) and azimuth ($\phi$), in degrees, of the selected Eigenmike microphone capsules. The radius of the microphone is 4 cm. The origin of space is the center of the Eigenmike.

In the following, we consider the whole dataset described in (Sec. 4.1). The results of the comparison are given in Fig. 3. As expected performance decrease as reverberation and scene complexity increase, regardless of the signal domain. However, in most configurations, separating the sources in the HOA domain resulted in better performance measures compared to the microphone domain. We can clearly see a gain of 7 to 12 dB for the least challenging sound source configuration, and a gain of 1 to 6 dB for the most challenging one. Tab. 3 summarizes the difference in SDR values between the HOA domain and the microphone domain for configurations (A) and (D): the SDR is almost always higher in the HOA domain, regardless of the reverberation or song. As well, the gap between the performance obtained in the two domains reduces as the complexity of the scenario increases, with a more prominent influence of reverberation time. Separating sources in the HOA domain results in a 1 to 12 dB increase in signal-to-distortion ratio, compared to the microphone domain.

| | $RT_{60}(s)$ | 0 | 0.2 | 0.4 | 0.7 |
|---|---|---|---|---|---|
| A | max | 21 | 14.35 | 11.9 | 12 |
| | median | 12.43 | 7.69 | 7 | 6.84 |
| | min | 4.3 | 2.52 | 2.5 | 2.9 |
| D | max | 10.17 | 6.6 | 6.7 | 6.32 |
| | median | 6.05 | 0.83 | 1.52 | 2.45 |
| | min | -1.84 | -6 | -5 | -4 |

**Table 3**: $\Delta SDR = SDR_{HOA} - SDR_{MIC}$, in dB, for scenarios A and D.

One reason may explain these results. Indeed, in FASST's EM algorithm, the empirical covariance matrix is inverted while estimating the first Wiener filter [13] and the numerical stability of this inversion differ in the two signal domains. We calculated the condition number of the empirical covariance matrix in both domains for a random example picked from the dataset. It appeared that, for frequencies below 2 kHz, the condition number was generally higher in the microphone domain than in the HOA domain, and could be about 1000 times greater for some frequency values. Therefore, the conversion of the microphone signals into HOA signals seems to act as a pre-conditioning for the EM algorithm.

Having established the interest of performing the source separation in the HOA domain with FASST, we now compare it with the reference methods. Results are presented in Fig. 4. FASST clearly outperforms the reference methods. This is because, contrary to the reference methods which are solely based on spatial cues, FASST also exploits spectral cues. This gives FASST an advantage when sources are close to each other and spatial information is more ambiguous. Although this fact has already been observed in microphone domain source separation [21], we confirm it here also on HOA-domain source separation.

Surprisingly, FASST outperforms the PIV method even in anechoic environment where the PIV method could have been expected to give the best results in terms of performance. Indeed, 9 signals should be enough to form a beam toward one source and cancel 3 interfering sources at the same time. This can be explained with the fact that encoded HOA signals don't match perfectly the theoretical signals. This imperfection is mainly caused by the physical limitations of the microphone array. Indeed, the capsules of the Eigenmike are relatively close to each other, which results in spatial aliasing and a loss of lower frequencies [22].

## 6. CONCLUSION

In this paper we investigated for the first time the ability of the local Gaussian model to handle the source separation problem in the HOA domain. To this aim we have established the model's equations in the HOA domain and run numerical experiments. Our simulation results show that applying a local Gaussian model-based source separation method in the HOA domain typically results in the SDR increasing by 1 to 12 dB, compared to the microphone domain with the same number of microphones/channels $I = M = 9$, including in challenging situations such as reverberant environments and complex source configurations. In future work we will explore using proximity microphones in order to guide the source separation and improve its performance, and finally employ this method to allow navigation through HOA sound scenes.

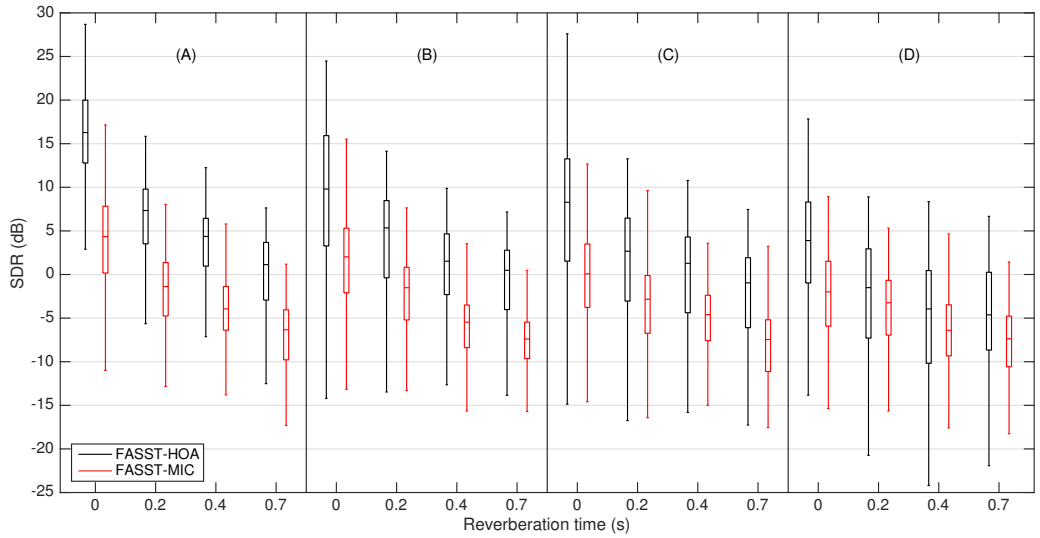


**Fig. 3**: Comparing FASST's performance in the HOA domain to FASST's performance in the microphone domain, $I = M = 9$

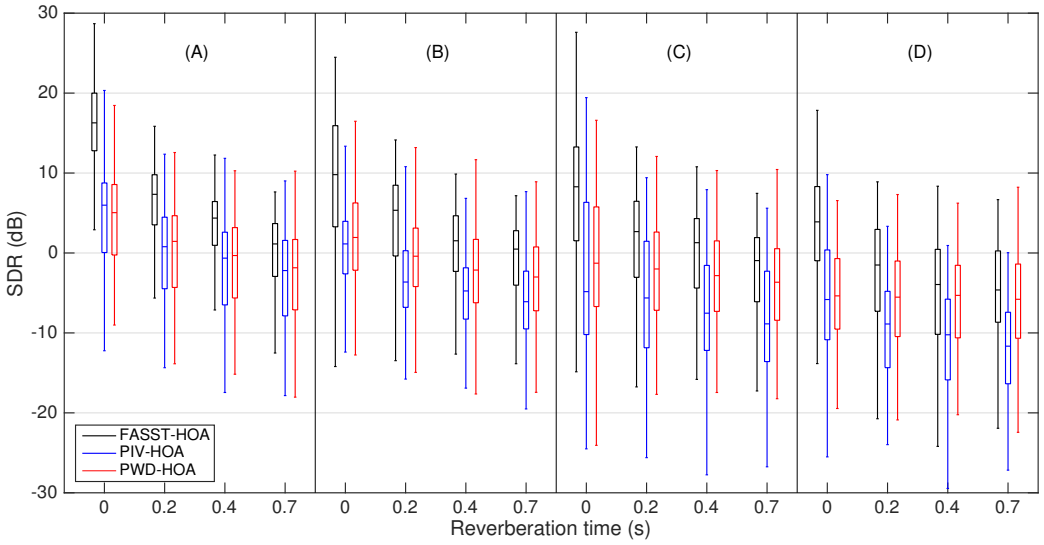


**Fig. 4**: Comparing FASST to the reference methods in the HOA domain.

## 7. REFERENCES


[1] Michael A Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[2] Michael A Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[3] Jérôme Daniel, Jean-Bernard Rault, and Jean-Dominique Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *Audio Engineering Society Convention 105*. Audio Egineering Society, 1998.

[4] Jérôme Daniel, Sebastien Moreau, and Rozenn Nicol, "Further investigations of High-Order Ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

[5] Alex Southern, Jeremy Wells, and Damian Murphy, "Rendering walk-through auralisations using wave-based acoustical models," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 715–719.

[6] Joseph G Tylka and Edgar Choueiri, "Soundfield navigation using an array of Higher-Order Ambisonics microphones," in *Audio Engineering Society conference: 2016 AES Interna-*

*tional Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[7] Matthias Kronlachner and Franz Zotter, "Spatial transformations for the enhancement of ambisonic recordings," in *International Conference on Spatial Audio*, 2014.

[8] Andrew Allen and W. Bastiaan Kleijn, "Ambisonic soundfield navigation using directional decomposition and path distance estimation," in *International Conference on Spatial Audio*. Graz, 2017.

[9] Mohammed Hafsati, Nicolas Epain, and Jérôme Daniel, "Editing ambisonic sound scenes," in *International Conference on Spatial Audio*. Graz, 2017.

[10] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[11] Alexey Ozerov and Cédric Févotte, "Multichannel Nonnegative Matrix Factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[12] Emmanuel Vincent, Shoko Araki, and Pau Bofill, "The 2008 Signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 734–741.

[13] Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraa-Labastie, Xabier Jaureguiberry, Dung T Tran, and Frédéric Bimbot, "The Flexible Audio Source Separation Toolbox Version 2.0," in *ICASSP*, 2014.

[14] Emmanuel Vincent, Simon Arberet, and Rémi Gribonval, "Underdetermined instantaneous audio source separation via Local Gaussian Modeling," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 775–782.

[15] Rozenn Nicol, "Sound spatialization by Higher Order Ambisonics: Encoding and decoding a sound scene in practice from a theoretical point of view," in *International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[16] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[17] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*. Citeseer, 2010, pp. 1–6.

[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "BSS_EVAL toolbox user guide–revision 2.0," 2005.

[20] MH Acoustics, "EM32 Eigenmike microphone array release notes (v17. 0)," *25 Summit Ave, Summit, NJ 07901, USA*, 2013.

[21] Emmanuel Vincent, *Contributions To Audio Source Separation And Content Description*, Ph.D. thesis, Université Rennes 1, 2012.

[22] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet, "3D sound field recording with higher order ambisonics–Objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, 2006, pp. 20–23.