Unsupervised environmental sound source localization and acoustic image analysis of geometry-optimized spherical microphone arrays using the generalized cross-correlation

Lucas Henrique Teixeira Carneiro, Alain Berry

UNSUPERVISED ENVIRONMENTAL SOUND SOURCE LOCALIZATION AND ACOUSTIC IMAGE ANALYSIS OF GEOMETRY-OPTIMIZED SPHERICAL MICROPHONE ARRAYS USING THE GENERALIZED CROSS-CORRELATION

Lucas Carneiro[1,2] **Alain Berry**[1,2]
[1] Groupe d'Acoustique de l'Université de Sherbrooke, Université de Sherbrooke,
2500 Boulevard de l'Université, Sherbrooke, QC, Canada J1K 2R1
[2] CIRMMT, McGill University,
527 Rue Sherbrooke Ouest, Montréal, QC, Canada H3A 1E3
lucas.carneiro@usherbrooke.ca, alain.berry@usherbrooke.ca

ABSTRACT

Spherical microphone array processing has been deployed as a tool for sound source localization and diagnosis using acoustic image techniques via algorithms such as the generalized cross-correlation (GCC). However, some acoustic image analysis tools used in this domain remain primitive. In this work, it is proposed a method for unsupervised source localization in acoustic images using the image segmentation principle, a method to compute the main-lobe width (MLW) and the (maximum) side-lobes levels (SLL) that consider the spherical property of the beamforming projection and a method to compute the source localization uncertainty according to a covariance ellipse estimation that uses image gradients, the MLW and the SLL. These methods are numerically tested under different acoustic conditions and with optimized microphone arrays.

1. INTRODUCTION

In many applications, it is desirable to understand the location of one or more sound sources and their propagation properties. Such a diagnosis can be latter used for the hierarchical classification of the sources as contributors to the overall environment sound level. The classification is an important step prior to the design of noise abatement and control strategies. Microphone array processing can be used to generate acoustic images that serve for this purpose [1–5]. What is described next is a method to efficiently generate, process and extract information of these images.

2. METHOD

2.1 Generalized cross-correlation - Phase alignment (GCC-PHAT)

The generalized cross-correlation (GCC) between microphone signals may be computed with an IFFT:

$$R_{m_i,m_{ii}}(\tau) = \sum_{f=0}^{k-1} \Psi(f) C_{X_{m_i},X_{m_{ii}}}(f) e^{\frac{j2\pi f\tau}{k}}, \qquad (1)$$

where f the frequency index, k is the sampling frequency, $\tau = (\Delta t_{m_{ii}\mathbf{k}} - \Delta t_{m_i\mathbf{k}})$ is the (unknown) time-lag between microphone signals $X_m(\omega)$, $\Psi(f)$ is an inter-microphone weighting function, $C_{X_{m_i},X_{m_{ii}}}(f)$ is the power cross-spectrum of the signals and $\mathbf{k}$ is a steering vector. Equation 1 is a likelihood estimator for the cross-correlation function proposed by Knapp [6] and assumes the stationary nature of the signals during the observation time. The argument τ maximizing the function provides the estimation of the time-delay corresponding to the direction of sound arrival for the pair of microphones considered.

Since $\tau_{m_i,m_{ii}} = -\tau_{m_{ii},m_i}$ and $R_{m_i,m_{ii}}(\tau) = R_{m_{ii},m_i}(\tau)$, the non-redundant beamformer for the GCC is written as:

$$y^2(\mathbf{k}) = \sum_{i=1}^{M} \sum_{ii=i+1}^{M} R_{m_i,m_{ii}}(\tau). \qquad (2)$$

Because $R_{m_i,m_{ii}}(\tau)$ is a time-observation function, certain time-lags might surpass the maximum physically allowable time-lag $max(\tau_{m_i,m_{ii}})$ between all pairs of microphones. Therefore, the cross-correlation computed in Equation 1 needs to be truncated in a set of feasible time-lags such that $\tau \leq max(\tau_{m_i,m_{ii}})$ before further processing [1].

Hence the beamformer is obtained from the superimposition, for each pair of microphones, of cross-correlation values obtained from the interpolation of truncated cross-correlation values $\check{R}_{m_i,m_{ii}}$ into the set of time-lags belonging to the scanning region of $\mathbf{k}$. The weighting function $\Psi(\omega)$ may be set to 1 or an adaptive beamformer created using the phase-alignment filter (PHAT):

$$\Psi_{PHAT_{m_i,m_{ii}}}(\omega) = \frac{1}{|X_{m_i}(\omega) X^*{}_{m_{ii}}(\omega)|}, \qquad (3)$$

that was mathematically proven to give, when the number of observations is sufficiently large and outnumbers the number of sources, an optimal localization in a maximum likelihood sense for many acoustic conditions [7]. PHAT is therefore an interesting technique to be used with array systems. However, it distorts the level information of the final beamformer.

2.2 Acoustic image generation

In the far-field, the time-lag may be computed as a function of the scanning vector as:

$$\Delta t_{m\mathbf{k}} = \mathbf{r}_m \cdot \mathbf{k}/c_0, \tag{4}$$

where $\mathbf{r}_m$ is the microphone position vector and c_0 is the speed of the sound. Fig. 1 demonstrates the realization of a spherical scanning region using spherical microphone arrays. The virtual projection surface is defined at a distance $\|\mathbf{f}\|$ of the microphone array and with angular resolution $(\Delta\theta, \Delta\varphi)$ defined by an equi-rectangular grid.

A truncated cross-correlation $\check{R}_{m_i,m_{ii}}$ is equivalent to the spatial likelihood function (SLF), the zone in the space where all time-lags for a given pair of microphones are the same and the probability to find a source using time-lag localization methods is maximum [2]. The 3D geometric locus of the SLF is an hyperboloid, or an hyperbola when projected on a surface.

The superimposition of cross-correlation values are calculated on the spherical surface of Fig. 1 using arithmetic (AM) or geometric means (GM), respectively [3]:

$$\begin{aligned} y^{2^{AM}}(\mathbf{k}) &= \frac{1}{P}\sum_{i=1}^{P}\check{R}_{p_i}(\tau), \\ y^{2^{GM}}(\mathbf{k}) &= \prod_{i=1}^{P}\left|\check{R}_{p_i}(\tau)\right|^{\frac{1}{P}}, \end{aligned} \tag{5}$$

where P is the number of pairs of microphones. The maximum of the superimposition is the common region where many SLFs' cross and indicates the sound source localization. The GM tends to attenuate background noise on the final image and suppress side-lobes. However, it distorts levels information on the final image and diminishes the level of weaker sources, making their localization more difficult.

The final image is obtained normalizing the beamformer response on the surface of projection and applying a logarithmic scale. Fig. 2 demonstrates the realization of this process for a circular microphone array of 4 microphones. The proposed acoustic conditions and signal processing parameters used in this computation are standard throughout this work unless otherwise stated. The image scale is homogeneous to $p^2_{RMS}(t)$.

2.3 Optimized microphone arrays

The GCC-PHAT is optimal in a maximum likelihood sense and should indicate a source localization on the maximum of a gaussian distribution on the acoustic image. However, biased SLFs' superimposition will ruin this optimality and amplify side-lobes generation. In order to reduce biased superimposition and induce a phenomenon called 'spatial-whitening' on the acoustic image, Carneiro [4] proposed the following geometric criteria that must be maximized:

$$\begin{aligned} \Gamma &= \frac{\sum_{i=1}^{P}\left|\gamma\left(\mathbf{k}_{p_i}, \mathbf{k}'_{p_i}\right)\right|}{P}, \\ D &= \frac{\sum_{i=1}^{P}\left|d_{p_i} - d'_{p_i}\right|}{P}, \end{aligned} \tag{6}$$

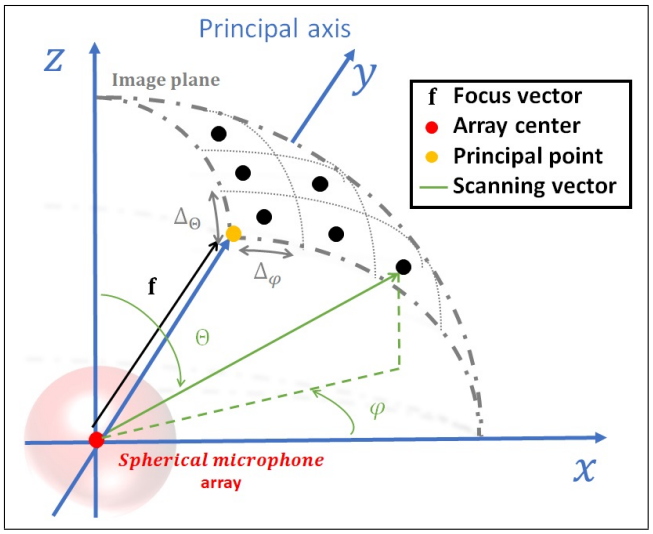

Figure 1. Virtual spherical surface of projection for the beamforming: $x = \|\mathbf{f}\|\sin(\theta)\cos(\varphi)$, $y = \|\mathbf{f}\|\sin(\theta)\sin(\varphi)$ and $z = \|\mathbf{f}\|\cos(\theta)$. $\Delta\theta = \Delta\varphi = 1°$ in this work.

referred to as the mean orientation difference (the mean of the sum of the difference of orientation between pairs of microphones) and the mean separation difference (the sum of the difference of separation between pairs of microphones), respectively. P is the number of microphones, γ is the aperture angle between the unit orientation vector $\mathbf{k}_p$ of a given pair of microphones and the best alignment vector $\mathbf{k}'_p$ (smallest angle) among any other pair to $\mathbf{k}_p$, d_p is the linear separation of a given pair of microphone and d'_p is the closest separation among any other pair to d_p. $\mathbf{k}_p$ and d_p are computed as follows:

$$\begin{aligned} \mathbf{k}_{p_{i,ii}} &= \frac{\mathbf{r}_{m_i} - \mathbf{r}_{m_{ii}}}{\|\mathbf{r}_{m_i} - \mathbf{r}_{m_{ii}}\|}, \\ d_{p_{i,ii}} &= \|\mathbf{r}_{m_i} - \mathbf{r}_{m_{ii}}\|, \end{aligned} \tag{7}$$

where i, ii are indices for the microphones of the pair and $\mathbf{r}_m$ the vector coordinates of the microphone.

The optimization of a spherical microphone array of M microphones, diameter L and eventually pre-defined linear and/or non-linear design constraints requires the maximization of one or both geometric criteria using a single or multi-objective genetic algorithm (GA) [8].

2.4 Image segmentation and unsupervised localization

The topology of a typical acoustic image may be generalized as a series of local peaks (maximum) indicating the localization of multiple sources. These areas will be called foreground. Side-lobes may appear as local peaks on the foreground. The background of the image may present a noisy allure with many local peaks and valleys separated by a, sometimes spatially variable, dynamic range from the foreground.

A multiple sound source localization algorithm may be designed to localize the sources on a segmented image. A segmented image is a binarized version of the original im-

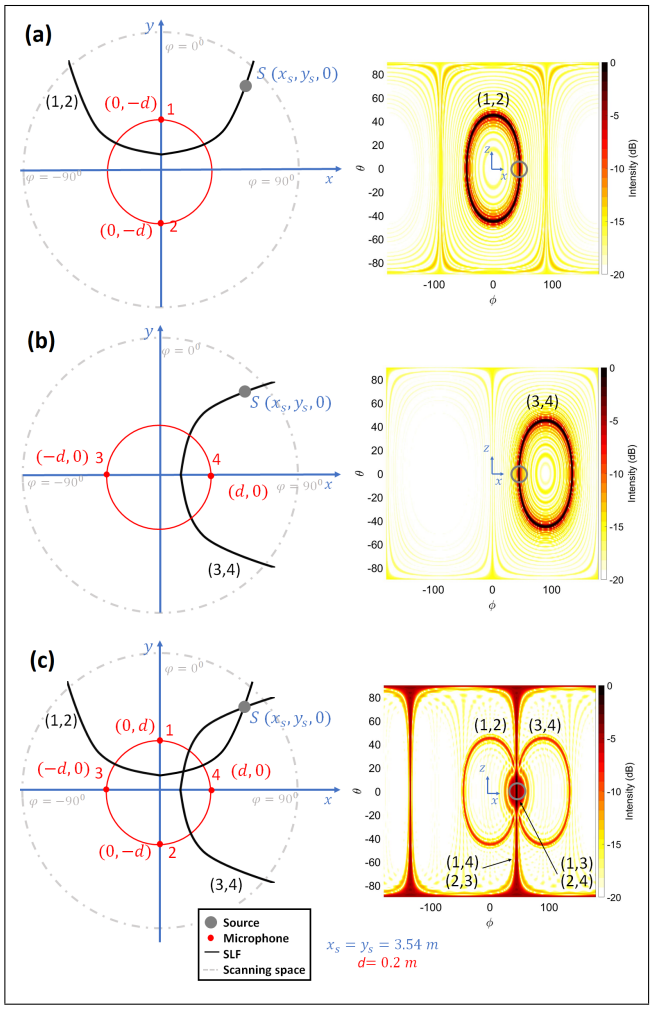

Figure 2. A wide-band monopole is considered and modeled using the free-field radiation equation of a pulsating sphere with low additive noise (15 dB environmental and 64 dB at the microphones). The GCC-PHAT-AM algorithm is used sampling 1 s of time signal at 32768 Hz. (a)/(b) SLF for a pair of microphones projected on the $x-y$ plane and on the steering space. (c) The superimposition of SLFs generates a pin-point localization (in the acoustic image, SLFs for all microphone pairs are depicted).

age with foreground indicated as 0 (black) and the background indicated as 1 (white). Assuming that segmented black areas for the different sources are not connected, the algorithm uses the binary image to locate the areas in black. It transfers this localization information to the original acoustic image, where the position of the local maximum for each area is extracted and attributed as a new sources. If the segmented areas or two sources are connected, one of the sources will not be located, resulting in a false-negative. If a side-lobe is segmented as foreground (black), it will result in a false-positive.

Three segmentation methods suitable to acoustic images are discussed in this paper: the h-minima morphological filter [9], Otsus' method [10] and Bradleys' method [11]. While computer graphics applications for these methods are widely discussed in the literature, usage in acoustic images seems to be new up to the authors' knowledge. It is therefore important to understand the principle of these methods, their hypotheses and the necessary conditions in order to achieve good segmentation results.

The h-minima filter can be used to extract a series of extreme values from the image using a contrast criterion. It suppresses all minima whose deepness relative to a user-defined vicinity is smaller or equal to a threshold h (a non-negative scalar), performing a reconstruction by erosion of y from $y + h$, as follows:

$$HMIN_h(y) = R_y^{\epsilon}(y + h), \tag{8}$$

where R_y^{ϵ} is an erosion transformation of the grid value y. Intuitively, the filter may be compared to the geological erosion mechanism: the erosion residue of the highest ridges (global maxima) covers the minor ones (local minima). The segmented image presents a foreground with minor modifications (the value at the peak is the last to be eroded) while the background is flattened.

Finally, the resulting image may be binarized. Nonetheless, the h-minima filter is a supervised segmentation strategy because the deepness threshold h needs to be set beforehand and only allows supervised source localization.

Otsus' and Bradleys' methods are both non-parametric and unsupervised automatic image segmentation methods based on discriminant and integrative analysis, respectively. The acoustic image needs to be converted to a gray-scale of N pixels and L levels from 0 to 1 (0 is black and corresponds to the strongest source and 1 is white).

Otsus' method evaluates and selects an optimal threshold(s) that maximizes the separability between two or more classes from the gray-level histogram of the image. According to the discriminant analysis, the optimal threshold(s) maximizing the separability of the classes is/are the one(s) maximizing the between-class variance computed from the gray-level histogram. For instance, segmentation is achieved defining two classes: C_0 is the set of pixels with levels up to k (the foreground) and C_1 is the set of pixels with levels above k (the background), with k being the optimal threshold separating both classes. Once k is estimated, it can be used to binarize the original acoustic image.

The discriminant analysis assumes that the gray-level histogram has bimodal distribution. This condition is ideally secured when the surface distribution of the class C_0 on the image is of the same order of C_1, when the mean level difference between C_0 and C_1 is large and when the variance of each class is small [12]. In other words, the image is not extensively corrupted by noise and the background is not affected by 'non-uniform illumination', a condition where the background of the image is affected by a non-uniform dynamic range.

On acoustic images, if the main lobe and the rest of the image are considered to be distinct classes, the above conditions are better met when the side-lobe levels (SLL) are smaller and spatially constant and the main-lobes width (MLW) are larger.

Bradleys' method computes a moving average on a window of $n \times n$ pixels (also known as kernel) on the integral image of the gray-scale. If the value of the current pixel is smaller than t percent the average of the moving average the pixel is set to black (0), otherwise it is set to white (1). t is the sensitivity of the segmentation operation. Under the hypothesis that the image contains dominant background pixels (white) and that the foreground is distributed, this operation preserves hard contrasts and ignores soft gradient changes even under 'non-uniform illumination' conditions. The kernel needs to be larger than the foreground components or may lead to miss-classification. If the kernel is too large it may lead to a loss on the ability to segment fine details. If t is too low (higher sensitivity), the segmentation becomes too sensitive to noise. Too high and sources are not detected.

It is important to note that the segmentation methods do not need to operate on the logarithmic version of the acoustic image. They may also operate on the linear version and with any other beamforming technique. Ideally, the image needs to be able to efficiently suppress side-lobes, which can be accomplished using optimized microphone arrays.

2.5 Metrics extraction

The standard used in this work to compute the main-lobe width (MLW) and the (maximum) side-lobe level (SLL) is similar to that presented by Christensen [5] and is synthesized in Fig. 3.

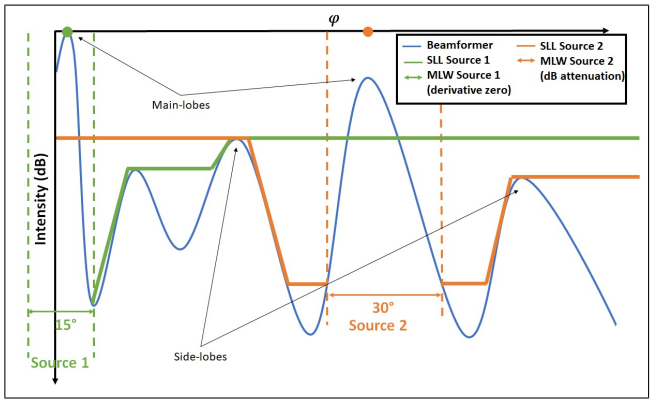

Figure 3. Definition for main-lobe width (MLW) and (maximum) side-lobe level (SLL) using a fictive 1D beamformer with two located sources.

The MLW is a metric used to evaluate the size of the main-lobe and is in most cases an angular value. It is generally defined as the angular aperture between the first two neighboring minima (derivative zero) on the localization topology or two points defined by a certain dB attenuation (usually up to -10 dB).

The SLL is a profile obtained excluding the MLW for all located sources. It keeps track of the maximum topology level out of the MLW. It can be simply characterized by the maximum tracked level.

Fig. 3 is a 1D topology for which metrics can be easily computed in the direction obtained from the abscissa. In the case of 2D acoustic images, it is not realistic to compute these metrics for the orthogonal axes (θ,φ). A more precise computation can be executed in the radial directions of the source localization but requires an interpolation of the topology originally generated from an equi-rectangular grid to a polar grid centered on the source position (Fig. 4). This is accomplished positioning a spherical

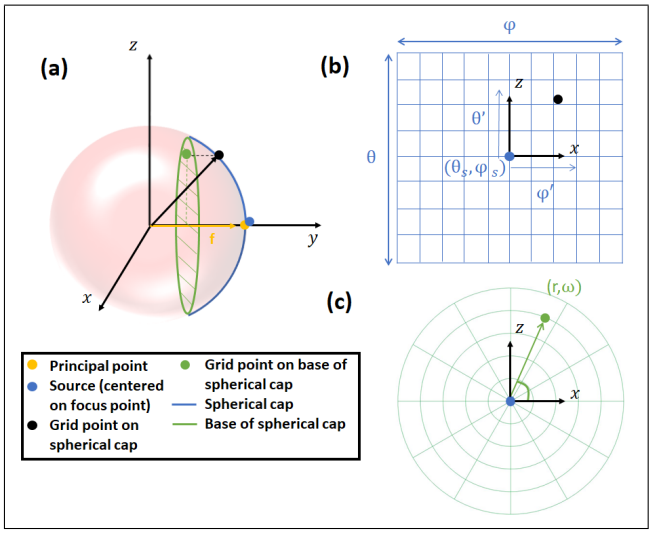

Figure 4. (a) Equi-rectangular grid points can be transferred from the spherical cap to its base (green plane): $r_z = \|\mathbf{f}\| \sin(\theta')$, $r_x = \|\mathbf{f}\| \sin(\varphi')$. This transformation introduces large pole distortion. Although the distortion is minimal for any points close to the focus point. (b) Representation of an equi-rectangular grid centered on the source position. (c) Representation of a polar grid centered on the source position.

cap centered on the sound source location (Fig. 4 (a)) and converting the angular positions on the equi-rectangular grid (Fig. 4 (b)) to the base of the spherical cap. These samples can be finally used to interpolate a regular polar grid (Fig. 4 (c)). The standard polar grid discretization in this work consisted of 360 radially distributed points with constant spacing of 1° and 100 linearly spaced points in the polar axis of a hemi-sphere cap.

Once the local equi-rectangular topology is interpolated to a polar grid, the MLW may be computed radially using the appropriate criterion (derivative zero or dB attenuation) and smoothing the data to compensate the noise allure of the slope: a radius r_i is obtained for each elements distributed on the angular coordinate of the polar grid. This radius may be converted back to a radial aperture angle on the equi-rectangular spherical grid as $\gamma_i = \arcsin(r_i / \|\mathbf{f}\|)$. The set of radial aperture angles form a contour. The MLW may be represented as a mean radial aperture angle that fits a circular spherical cap on the MLW contour:

$$MLW_\gamma = \arccos(1 - 2S_{MLW}) = \arccos\left(1 - \frac{4\pi \|\mathbf{f}\|^2}{K} \sum_{i=1}^{K} (1 - \cos(\gamma_i))\right), \quad (9)$$

or as a fraction of the equi-rectangular spherical grid sur-

face:

$$MLW_{\%} = \frac{1}{K}\sum_{i=1}^{K}(1-\cos(\gamma_i)), \tag{10}$$

where S_{MLW} is the area computed from the angular aperture criteria for the MLW and K denotes the number of elements distributed on the angular coordinate of the polar grid.

The global SLL can be computed in the sequence. Its maximum is the maximum value of the acoustic image topology on the equi-rectangular grid excluding the MLW contour for all identified sources, in dB. Its mean is the mean of the acoustic image topology excluding the MLW contour for all identified sources, in dB. A SLL may be computed individually for each identified source. On a polar grid, the maximum topology level can be tracked along each radial angle ω_i. A SLL image in polar coordinates is obtained. Based on this image, maximum and mean SLL may be computed as described in the above paragraph.

Assuming that the main-lobe follows a bi-variate normal distribution in $\theta-\varphi$, a confidence interval centered at the peak of this distribution can be estimated in the form a covariance ellipse with orthogonal maximum and minimum variations. Modifying the method discussed in Brooks [13], who proposes a general covariance estimation method for vision systems, the principal axes and orientation of the ellipse can be computed from the eigenvalue decomposition of the following positive semi-definite (hence symmetric for real numbers) covariance matrix:

$$\mathbf{\Sigma}_{\theta,\varphi}(\theta_s,\varphi_s) = \frac{1}{\widehat{E^2}_\theta\widehat{E^2}_\varphi - \widehat{E\theta E}_\varphi\widehat{E\varphi E}_\theta}\begin{bmatrix}\widehat{E^2}_\theta & \widehat{E\varphi E}_\theta \\ \widehat{E\theta E}_\varphi & \widehat{E^2}_\varphi\end{bmatrix}, \tag{11}$$

where $\widehat{E}$ denotes the directional variance of a locally normalized distribution computed for a certain source located in (θ_s,φ_s). The directional variance is computed, for instance, in the θ direction of an equi-rectangular spherical mesh as:

$$\widehat{E_\theta}(\theta_s,\varphi_s) = \sum_{(i,ii)\in MLW_s}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\gamma^2\left(\mathbf{k}_{i\Delta_\theta,ii\Delta_\varphi},\mathbf{k}_{\theta_s,\varphi_s}\right)}{2\sigma^2}} \times\mathbf{E}_\theta(\theta_s+i,\varphi_s+ii), \tag{12}$$

and is analogous in the φ direction. γ is the aperture angle between two unitary orientation vectors $\mathbf{k}$ computed using grid coordinates. The computation is done on the zone interior to the MLW contour (derivative zero) with i and ii being grid counters representing grid coordinates angles in θ and φ, respectively. σ is the angular standard deviation and can be estimated as $1/3$ of the MLW_γ (derivative zero) (Equation 9). $\mathbf{E}_\theta$ is the gradient matrix of $\mathbf{E}$ in θ. $\mathbf{E}$ is the beamformer response (image value) in $\theta-\varphi$ normalized between 0 (lowest valley) and 1 (highest peak). The gradient captures the concept that the source is best and worst located in the directions of maximum and minimum changes on the acoustic image, respectively. The size of the ellipse is proportional to the MLW_γ (derivative zero), which defines the standard deviation, and the MLW_γ (dB attenuation). The MLW_γ (dB attenuation) is inversely proportional to the gradient, or inversely proportional to the source level and directly proportional to the SLL.

3. NUMERICAL RESULTS

3.1 Definition of microphone array geometries

It was decided that the microphone array should have $M = 18$ microphones and $L = 0.4$ m in diameter. A regular and a non-optimized random microphone distribution were defined as basis of comparison. Then, single and multi-objective optimization routines using genetic algorithms were set in order to obtain the best possible microphone distribution on the defined sphere according to the geometric criteria.

Also, for the experimental work, a spherical microphone array support of 15 cm in diameter had to be used: the 18 microphones are mounted on aluminum rods of approximately 7.5 cm in length, which can be screwed to 252 different threaded hole positions on the support surface. In order to optimize the microphones distribution in this configuration, single and multi-objective integer/discrete genetic algorithm optimization routines were coded. At each GA generation, the code considers a continuous distribution from 1 to 252 that maps the coordinates of the threaded holes of the support, for computational purposes. Then, it performs an unbiased truncation of this continuous mapping to integer numbers. This coding technique allows multi-objective integer/discrete optimization using GA, which is not available in most commercial algorithms.

In total, 9 microphone arrays geometries were defined. Their nomenclature, distribution and characterization according to the geometric criteria are depicted in Fig. 5.

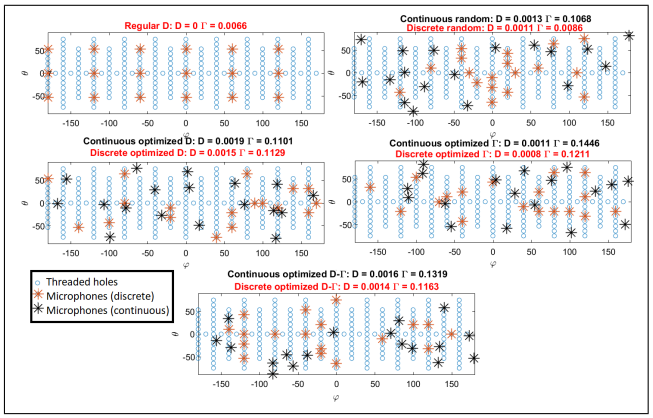

Figure 5. Microphone array geometry distributions.

For the single-objective optimization, the mean real criteria improvement varied from 15% (discrete Γ) to 52% (continuous Γ) with respect to the random arrays. For the multi-objective optimization, it provided improvements on the order of $35-10\%$ for $D-\Gamma$, respectively, in the discrete case and 30% for both $D-\Gamma$ in the continuous case.

3.2 Implementation of the segmentation and localization methods

For better understanding and first validation of the segmentation and localization methods, the solutions proposed in 2.4 are tested on an image created from 4 randomly distributed, wide-band and uncorrelated monopoles of identical levels and distances to microphone array 'continuous optimized D'. AM and GM acoustic images are tested with gray-scale images obtained from linear and logarithmic versions of the original acoustic image. Results are depicted on Fig. 6 and Fig. 7.

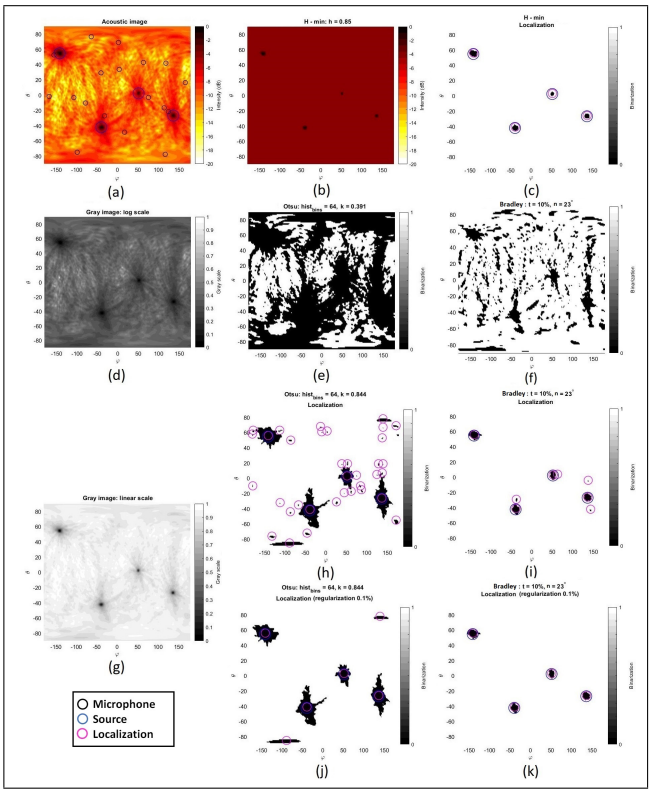

Figure 6. Source localization using different segmentation strategies and GCC-PHAT-AM images. (a) GCC-PHAT-AM. (b) Segmented image using h - minima transform. (c) Localization obtained from binarization of (b). (d) Image (a) in gray-scale. (e) Otsu segmentation of image (d). (f) Bradley segmentation of image (d). (g) Linear version of image (a) in gray-scale. (h) Otsu segmentation of image (g) and localization. (i) Bradley segmentation of image (h) and localization. (j) Regularized Otsus' segmentation of image (g) and localization. (k) Regularized Bradleys' segmentation of image (g) and localization.

As shown in Fig. 6 (c), a supervised method successfully segments and locates all sources if the threshold h is correctly estimated by the user. On the other hand, an unsupervised method may be deployed segmenting gray images (Fig. 6 (e) and (g)) into binary images (Fig. 6 (e-f), (h-i) and (j-k)).

In the case of GCC-PHAT-AM images, segmentation and localization are poorly achieved in log-scale gray images with Otsus' method: in Fig. 6 (e), the poor segmentation is due to the small and spatially variant level differences between classes (large and variable SLL causing 'non-uniform illumination') on the image. More generally, the logarithmic operation reduces the mean level difference between the classes on the image (they present larger SLL) and the variance of each class becomes too high due to side-lobes and noisy (irregular) background. Unde these conditions, the image does not present a bimodal distribution.

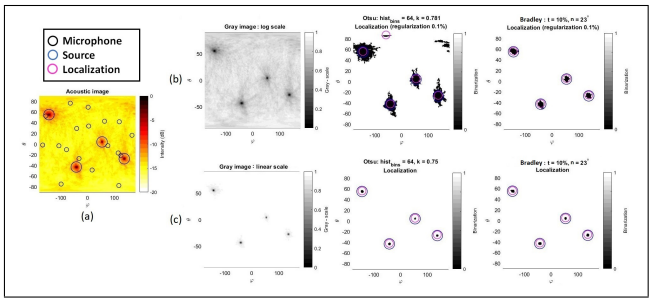

Figure 7. Source localization using different segmentation strategies and GCC-PHAT-GM images. (a) Acoustic image. (b) Segmentation obtained from log-scale image with regularization and localization. (c) Segmentation obtained from linear-scale image without regularization and localization.

On the other hand, segmentation and localization via Otsus' method are more effective on linear-scale gray images and may be further improved with regularization. Regularization is a common technique in image segmentation and consists in whitening very small dark connected areas of the segmented image for which surfaces are smaller or equal than a certain percentage of the full image surface. It is proposed a regularization of 0.1%, for which the localization is optimal without risk of missing the localization of potential sources (false-negatives).

In the case of GCC-PHAT-GM images, image segmentation via Otsus' method is also more effective because the GM operation reduces side-lobes and noisy background. Classes variances are reduced. On linear-scale GM images the variances are sufficiently small and the quality of the segmentation is good enough so that regularization is not necessary: in Fig. 7 (b), only 1 false-positive is obtained.

The Bradleys' method is deployed using a kernel size $n = 23°$ and sensitivity $t = 10\%$, values that are of standard usage in the literature. In practice, the size of the kernel is a conservative estimation of the MLW so that the source can be fitted inside the kernel. This choice reduces the sensitivity of the segmentation to noise.

Even under the presence of 'non-uniform illumination' due to side-lobes, segmentation and localization are in general successful on GCC-PHAT-GM and GCC-PHAT-AM linear-scale images with Bradleys' method: the segmentation is successful in Fig. 6 (k) and Fig. 7. 3 false-positives are detected in Fig. 6 (i). However, the localization can be easily improved with regularization.

The superior performance can be explained with two arguments: first, the images contain dominant background

pixels and distributed foreground and second, the level difference between foreground and side-lobes is sufficiently high so that the images can be successfully segmented. The last point is not true for GCC-PHAT-AM log-scale images (Fig. 6 (f)): the image cannot be properly segmented using conventional settings because of the noisy background. In such cases a better performance may be achieved reducing the sensitivity (increasing t) at the expense, for a certain threshold, of a higher probability of false-negatives detection. A second but not recommendable solution would be to increase the amount of regularization on the image.

3.3 Metrics extraction of a monopole, a dipole and a vibrating panel on the center of the image

Three numerical experiments with sources radiating on a free-field, centered and 5 m from the array are carried in order to demonstrate the ability of the proposed metrics to diagnose point from extended sources or omnidirectional from directional sources. The metrics can be automatically computed once the the source localization is obtained using unsupervised or supervised localization methods.

Fig. 8 is the acoustic image obtained from the radiation of a pulsating sphere. Fig. 9 is the acoustic image obtained from the radiation of two very close pulsating spheres in phase opposition and with the same level in the microphone array as in Fig. 8. This source presents an ideal dipole directivity pattern with maximum radiation attenuation in the direction orthogonal to the main radiation axis. The main radiation axis presents a pattern similar to a monopole and therefore is not here analyzed. These models were validated experimentally presenting good agreement with acoustic images obtained from a B&K 4295 omnidirectional source, baffled and unblaffed speakers. However, a real dipole obtained from an unblaffed speaker presents limited attenuation in the orthogonal direction and the presence of two sources as in Fig. 9 is not observed.

Fig. 10 is the acoustic image obtained from the radiation of a vibrating aluminum panel (0.6 m in azimuth, 0.55 m in elevation, 6 mm thickness and damping coefficient of 4%) facing the microphone array excited by an wide-band distributed force. It is obtained solving the vibroacoustic model of a panel and the Rayleigh integral equation in the frequency domain [14]. Vibrational and a acoustical parameters of this numerical model were validated with the literature [15]. Due to memory limitations, the sampling rate used in this analyze was Fs=8192 Hz. The panel presents a wide-band radiation pattern with a directivity pattern (evaluated in wide-band and using SPL) almost omnidirectional but with stronger radiation in the direction orthogonal to the panel. An acoustic image of a monopole with same level and processing parameters is obtained in Fig. 10 (e). The images are obtained using the GCC-AM because, due to the distributed nature of the source, it is more realistic to analyze metrics computed from a method that preserves images' absolute levels.

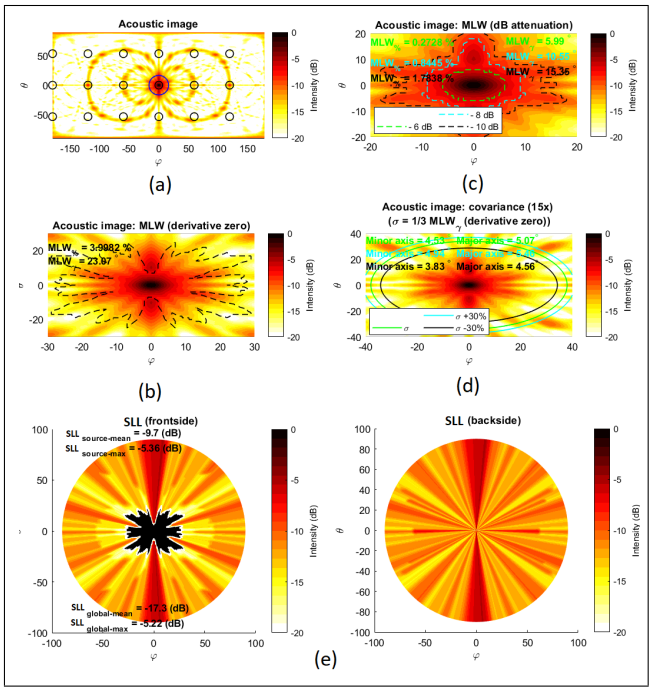

Figure 8. GCC-PHAT-AM image with a monopole on the center. (a) Acoustic image, microphones distribution ('Regular') and source location. (b) MLW (derivatize zero). (c) MLW (dB attenuation) for various attenuations. (d) Amplified covariance ellipse for various σ. (e) Front-side and back-side of SLL in polar plot.

4. CONCLUSION

It was proposed a method for unsupervised multiple sound source localization based on the segmentation principle of acoustic images. The best results are obtained with Bradleys' method and GCC-PHAT-GM. In the oral presentation we will support the affirmation that these methods work either numerically and experimentally with multiple sources at different levels and acoustic conditions. It will be discussed the positive impact of optimized microphone arrays on the method.

It was proposed a method to compute the MLW and the SLL that consider the spherical property of acoustic images and that is directionally invariant, i.e., it does not privilege user-defined analysis directions. It is demonstrated that the proposed metrics are capable to diagnose point from extended sources or omnidirectional from directional sources. In the oral presentation we will support the affirmation that optimized arrays tend to present larger MLW but smaller SLL, which is beneficial for segmentation and unsupervised localization.

Extensive results related to this research are available in the authors' thesis [16]. It is also part of the USPTO provisional patent application entitled 'System and method for tridimensional sound source diagnosis, localization and directivity reconstruction' and another conference paper [17].

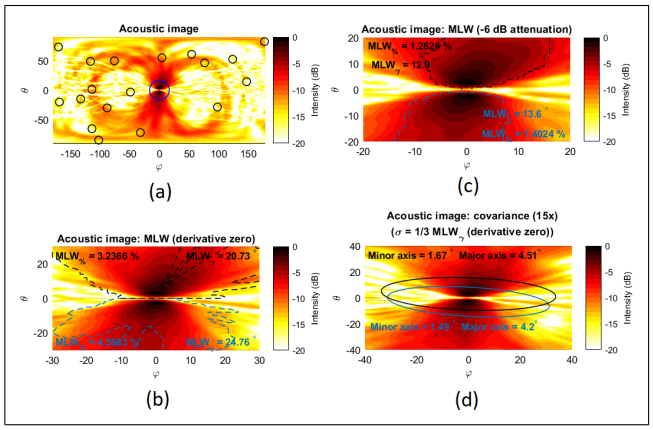

Figure 9. GCC-PHAT-AM image with a dipole radiating along elevation on the center. (a) Acoustic image, microphones distribution ('Continuous optimized D') and source location. (b) MLW (derivative zero). (c) MLW (dB attenuation). (d) Amplified covariance ellipses.

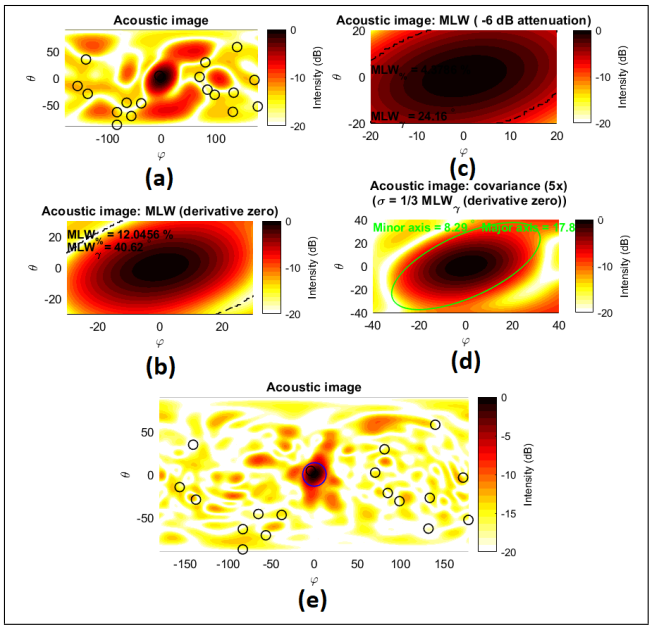

Figure 10. GCC-AM image with a panel or monopole on the center. (a) Acoustic image, microphones distribution ('Continuous optimized D-Γ') and source location. (b) MLW (derivative zero). (c) MLW (dB attenuation). (d) Amplified covariance ellipse. (e) Monopole. Metrics (not displayed): MLW_γ (derivative zero) = 30.7°, MLW_γ (dB attenuation) = 12.5°, major axis = 8.9°, minor axis = 5.0°.

5. REFERENCES

[1] e. a. T. Padois, "Acoustic source localization using a polyhedral microphone array and an improved generalized cross-correlation technique," *Journal of Sound and Vibration*, 2016.

[2] e. a. J. Velasco, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, 10/2012 2012.

[3] e. a. T. Padois, "On the use of geometric and harmonic means with the generalized cross-correlation in the time domain to improve noise source maps," *J. Acoust. Soc. Am.*, vol. 140, 2016.

[4] L. Carneiro, A. Berry, and T. Padois, "Optimization methodology of microphone arrays for environmental source localization using the generalized cross-correlation," BEBEC 2018, 2018.

[5] J. Christensen and J. Hald, "Beamforming," tech. rep., Brüel & Kjær, 2004.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, Aug 1976.

[7] e. a. C. Zhang, "Maximum likelihood sound source localization for multiple directional microphones," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, 2007.

[8] D. Kalyanmoy, *Introduction to Evolutionary Multiobjective Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[9] P. Soille, *Morphological Image Analysis: Principles and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2 ed., 2003.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, Jan 1979.

[11] D. D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *J. Graphics Tools*, vol. 12, 01 2007.

[12] e. a. S. Lee, "A comparative performance study of several global thresholding techniques for segmentation," *Computer Vision, Graphics, and Image Processing*, vol. 52, 08 1990.

[13] e. a. M. Brooks, "What value covariance information in estimating vision parameters?," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, 2001.

[14] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, vol. 34. 06 1989.

[15] A. Berry and J. Nicolas, "Structural acoustics and vibration behavior of complex panels," *Applied Acoustics*, vol. 43, no. 3, 1994. Structural Acoustics and Vibrations.

[16] L. Carneiro, *Tridimensional localization and directivity reconstruction of sound sources using the acoustic imaging structure from motion*. PhD thesis, Sherbrooke, 2020.

[17] L. Carneiro and A. Berry, "Sound source localization and diagnosis from multiple acoustic images," in *Internoise*, 2020.