Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

Mathematical Modelling and Optimization Strategies for Acoustic Source Localization in Reverberant Environments

PhD. Thesis Presented by

Jose FranciscoVelasco Cerpa

Advisors

Dr. Daniel Pizarro

Dr. Javier Macias-Guarasa

Alcalá de Henares, 2016

A quienes me quieren y a quienes quiero...

"It is not knowledge, but the act of learning,
not possession but the act of getting there,
which grants the greatest enjoyment."
Carl Friedrich Gauss

Acknowledgements

I would like to thank the labour of the external reviewers in revising this manuscript. Thank you for your time and for helping us to improve the present piece of work.

Thanks to Prof. Gonzalo Arce and Prof. Hervé Bourlard for hosting me during three months at the University of Delaware (USA) and IDIAP Research Institute (Switzerland), respectively.

Finally, I would like to thank all the people involved in this thesis: advisors and coauthors. Thank you for your help, suggestions and comments.

This work has been supported by the FPU Grants Program of the University of Alcalá, and has been also supported by the Spanish Ministry of Economy and Competitiveness under project SPACES- UAH (TIN2013-47630-C2-1-R), and by the University of Alcalá under projects DETECTOR and ARMIS.

Universidad de Alcalá

D. Manuel Mazo Quintas, Coordinador de la Comisión Académica del Programa de Doctorado en Electrónica: Sistemas Electrónicos Avanzados. Sistemas Inteligentes

INFORMA que la Tesis Doctoral titulada "Mathematical modelling and optimization strategies for acoustic source localization in reverberant environments", presentada por D. José Francisco Velasco Cerpa, bajo la dirección de los Doctores Daniel Pizarro Pérez y Javier Macías Guarasa, reúne los requisitos científicos de originalidad y rigor metodológicos para ser defendida ante un tribunal. Esta Comisión ha tenido también en cuenta la evaluación positiva anual del doctorando, habiendo obtenido las correspondientes competencias establecidas en el Programa.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Alcalá de Henares a 16 de Diciembre de 2016

Fdo.: Manuel Mazo Quintas

Universidad de Alcalá

D. Manuel Mazo Quintas, Coordinador de la Comisión Académica del Programa de Doctorado en Electrónica: Sistemas Electrónicos Avanzados. Sistemas Inteligentes

INFORMA que la Tesis Doctoral titulada "Mathematical modelling and optimization strategies for acoustic source localization in reverberant environments", presentada por D. José Francisco Velasco Cerpa, bajo la dirección de los Doctores Daniel Pizarro Pérez y Javier Macías Guarasa, reúne los requisitos exigidos por la Comisión Académica del Programa de Doctorado en "Electrónica: Sistemas Electrónicos Avanzados. Sistemas Inteligentes" para presentar dicha tesis por compendio de artículos.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Alcalá de Henares a 16 de Diciembre de 2016

Fdo.: Manuel Mazo Quintas

Universidad de Alcalá

D. Javier Macías Guarasa y D. Daniel Pizarro Pérez, directores de la Tesis Doctoral titulada "Mathematical modelling and optimization strategies for acoustic source localization in reverberant environments", presentada por D. José Francisco Velasco Cerpa

INFORMAN que dicha reúne los requisitos científicos de originalidad y rigor metodológicos para ser depositada y posteriormente defendida ante un tribunal.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Alcalá de Henares a 16 de Diciembre de 2016

Fdo.: Javier Macías Guarasa y Daniel Pizarro Pérez

Resumen

La presente Tesis se centra en el uso de técnicas modernas de optimización y de procesamiento de audio para la localización precisa y robusta de personas dentro de un entorno reverberante dotado con agrupaciones (arrays) de micrófonos. En esta tesis se han estudiado diversos aspectos de la localización sonora, incluyendo el modelado, la algoritmia, así como el calibrado previo que permite usar los algoritmos de localización incluso cuando la geometría de los sensores (micrófonos) es desconocida a priori.

Las técnicas existentes hasta ahora requerían de un número elevado de micrófonos para obtener una alta precisión en la localización. Sin embargo, durante esta tesis se ha desarrollado un nuevo método que permite una mejora de más del 30% en la precisión de la localización con un número reducido de micrófonos. La reducción en el número de micrófonos es importante ya que se traduce directamente en una disminución drástica del coste y en un aumento de la versatilidad del sistema final.

Adicionalmente, se ha realizado un estudio exhaustivo de los fenómenos que afectan al sistema de adquisición y procesado de la señal, con el objetivo de mejorar el modelo propuesto anteriormente. Dicho estudio profundiza en el conocimiento y modelado del filtrado PHAT (ampliamente utilizado en localización acústica) y de los aspectos que lo hacen especialmente adecuado para localización.

Fruto del anterior estudio, y en colaboración con investigadores del instituto IDIAP (Suiza), se ha desarrollado un sistema de auto-calibración de las posiciones de los micrófonos a partir del ruido difuso presente en una sala en silencio. Esta aportación relacionada con los métodos previos basados en la coherencia. Sin embargo es capaz de reducir el ruido atendiendo a parámetros físicos previamente conocidos (distancia máxima entre los micrófonos). Gracias a ello se consigue una mejor precisión utilizando un menor tiempo de computo.

El conocimiento de los efectos del filtro PHAT ha permitido crear un nuevo modelo que permite la representación 'sparse' del típico escenario de localización. Este tipo de representación se ha demostrado ser muy conveniente para localización, permitiendo un enfoque sencillo del caso en el que existen múltiples fuentes simultáneas.

La última aportación de esta tesis, es el de la caracterización de las Matrices TDOA (Time difference of arrival -Diferencia de tiempos de llegada, en castellano-). Este tipo

de matrices son especialmente útiles en audio pero no están limitadas a él. Además, este estudio transciende a la localización con sonido ya que propone métodos de reducción de ruido de las medias TDOA basados en una representación matricial 'low-rank', siendo útil, además de en localización, en técnicas tales como el beamforming o el autocalibrado.

Palabras clave: localización acústica; optimización y modelado matemático; entornos reverberantes; filtro PHAT; matrices TDOA.

Abstract

This thesis deals with the problem of indoor acoustic source localization using modern optimization strategies. It includes modeling, algorithms, and calibration, which allows using localization algorithms even when the geometry of the microphones is unknown. The aim of this thesis is to localize robustly and accurately speakers within a reverberant environment equipped with array of microphones.

The previous exiting techniques usually required a high number of microphones in order to get high accuracy. During this thesis, we have develop a new method which improves up to 30% the localization accuracy with a reduced number of microphones. Using a low number of microphones is important since it directly reduce the cost and improve the versatility of the final system.

On the other hand, we have performed a exhaustive analysis about the PHAT filtering (broadly used in acoustic localization), including all the phenomena involved in acquisition and signal processing. Our analysis improves the knowledge about PHAT filtering, modeling the main aspects involved in acoustic localization.

Previous model has yielded a sparse representation of the acoustic source localization scenario. This kind of representation has been demonstrated very convenient for localization since it allows to deal with multiple simultaneous sources easily.

Additionally, we have proposed a method for the calibration of pairwise distance using the diffuse noise present in a silent room. The new algorithm is related with previous methods based in coherence. Nevertheless, applying the developed model for PHAT filtering we have been able to introduce physical constraints based on the maximum expected distance between microphones. It allows to improve accuracy and reducing the computational cost.

Finally but not least, we have characterize TDOA matrices. We have propose several methods to robust denoise TDOA measurements exploiting low-rank properties of TDOA matrices. Therefore, these methods are not limited to acoustic source localization, but are useful for other techniques such as self-calibration and beamforming, and other technologies (e.g. radar, ultrasound).

Keywords: acoustic localization; mathematical modeling and optimization; reverberant environments; PHAT filtering; TDOA matrices.

Contents

Part I

Dissertation by compendium of publications: Dissertation summary

The present Thesis is presented in the format of compendium of articles regulated by the “regulations for elaboration, authorization and defense of the doctoral Thesis” of the University of Alcal? (UAH), approved by the UAH Governing board at 28th of September of 2016.

This first part of the manuscript gives an overview of the thesis work accordingly to the three indexed journal articles presented, an additional conference paper in a prestigious conference, and a technical report on the latest Thesis developments that have not been published yet. The first chapter provides a brief summary of the Thesis, it introduces the background of the studies, presenting a general revision of the state of the art, and describing and discussing the main contributions of the Thesis Work. In the second chapter, the summary and conclusions of the thesis are presented.

Chapter 1

Dissertation Summary

1.1 Introduction

Audio signals give very rich information as humans communicate mainly with speech. Nevertheless, is not totally clear yet how the human brain extract such information, so that, it is a topical matter. In this field, typical applications include speech recognition (including gender or emotion extraction), acoustic event recognition, source separation, localization, etc.

In particular, there is a considerable amount of publications focused on obtaining the exact position of any active acoustic source in a scene [1, 2]. Acoustic source localization has also a high impact in several domains, such as speech enhancement, beamforming techniques, and indoor human localization and tracking [3].

Localization of humans has a tremendous potential impact in diverse applied fields, opening new ways in how humans interact with machines. One important factor in indoor localization is the user awareness of the sensors used. Non-invasive technologies are preferred in this context, so that no electronic or passive devices are to be carried by humans for localization. The two non-invasive modalities that have been mainly used in indoor localization are those based on video systems and acoustic sensors.

This Thesis focuses on audio-based localization in a very general scenario, where unknown wide-band audio sources (e.g. human voice) are captured by a set of microphone arrays placed in known positions. The main objective of the Thesis is to use the signals captured by the microphone arrays to automatically obtain the position of the acoustic sources. This is an intricate problem due the existence of several source of error, such as periodicity in correlated signals, coherent noise or multi-path due to reverberation, which are inherent to the problem of acoustic source localization in indoor environments.

In a regular indoor scenario, most of the space is empty and there are only few sources active at the same time. This idea leads to a sparse representation of the localization problem and was the starting point of this Thesis [4]. Several methods have been proposed

to find the best sparse approximation of a linear system of equations, including brute force approaches as well as more computationally efficient approximate methods such as methods based on 'non linear programming' [5], and greedy pursuit [6–8]. Among all approximate solutions, l_1-norm based convex relaxations have flourished in the literature. It can be highlighted the Basis Pursuit Denoising method [9,10], originally introduced by [11] almost 40 years ago, but revisited with a profound theoretical study in the past decade, due to its intensive use in modern compressive sensing techniques [12,13]. These methods provide very effective polynomial time algorithms that, under certain circumstances, are even equivalent to the original l_0 based problems [10,13].

Nevertheless, the aforementioned methods need a descriptive model for the problem. This Thesis thus investigates mathematical models for the sound source localization problem, providing an analytical model which predicts the behaviour of both GCC-PHAT and SRP-PHAT [14]. We have also proposed several methods based on the previous model that improve the state-of-the-art in different tasks such as sound source localization [4] or microphone array pairwise distance based calibration [15]. Besides, we have proposed a novel denoising method for TDOA measurements robust to outliers, missing data and inspired by recent advances in robust low-rank estimation [16]. Furthermore, all methods proposed in this Thesis have been tested with data from real scenarios, such as the AV16.3 [17] dataset.

1.2 Previous Work

Existing approaches for acoustic source localization can be roughly divided into three categories [1,2]: Time delay based, beamforming based, and High-Resolution Spectral-Estimation based methods. This Thesis contributes to the first two categories.

1.2.1 Beamforming Based Methods

Beamforming based techniques [18] attempt to estimate the position of the source by maximizing or minimizing a spatial statistic associated with each position. For instance, in the Steered Response Power (SRP) approach, which is the simplest beamforming method, the statistic is based on the signal power received when the microphone array is steered in the direction of a specific location. Therefore, the position of the source is supposed to be consistent with the position corresponding to the maximum estimated signal power.

SRP-PHAT is a widely used algorithm for speaker localization based on beamforming. It was first proposed as such in [19][1], and is a beamforming based method which combines the robustness of the steered beamforming methods with the insensitivity to signal conditions afforded by the Phase Transform (PHAT). The classical delay-and-sum

[1] Although the formulation is virtually identical to the *Global Coherence Field* (GCF) described in [20].

beamformer used in SRP is replaced in SRP-PHAT by a filter-and-sum beamformer using PHAT filtering to weight the incoming signals.

The advantage of using PHAT is that no assumptions are made about the signal or room conditions [20], and this is the reason for the robustness of the SRP-PHAT method in reverberant scenarios, where the source is unknown. This method is usually defined as a standard for source localization, being a widely used algorithm for speaker localization. The main reasons of its popularity are its simplicity and robustness in reverberant and noisy environments, [21–25].

SRP-PHAT power map is a proper representation for acoustic source localization when only one source is active and many microphones are employed. Otherwise, the inherent redundancy of SRP-PHAT makes really hard to separate sources and to distinguish between sources and artifacts. In real scenarios, where only few sources are active at the same time and most of the space is empty, thus a sparse representation seems to be much more convenient.

1.2.2 TDOA Based Methods

These methods are based on estimating the time delay of signals relative to pairs of spatially separated microphones. In a second step, the time-difference of arrival information is combined with knowledge about the microphones' positions to generate a ML spatial estimator arising from hyperbolas intersected in some optimal sense [1, 2].

An accurate estimation of the time delay is essential for the good performance of this *time delay of arrival* (TDOA) methods. Assuming uncorrelated, stationary Gaussian signal and noise with known statistics and not multi-path, the maximum likelihood (ML) time-delay estimate is derived from a SNR-weighted version of the Generalized Cross Correlation (GCC) function [26]. Consequently, the two major sources of error in time delay estimation in real scenarios are coherent noise and multi-path due to reverberation. Some Different approaches have been proposed to deal with them.

A basic method consists in making the GCC function more robust, de-emphasizing the frequency-dependent weighting. The Phase Transform (PHAT) [26] is one example of this procedure which has received considerable attention as the basis of acoustic source localization systems due to its robustness in real world scenarios [14,27]. Other approaches are based in blind estimation of multi-path (room impulse response) [28] but they need a good initialization to perform well.

Other methods, which can be combined with PHAT, take advantage of the redundancy of the TDOA measurements from different pairs of microphones in order to reduce noise. In this regard, Gauss-Markov estimator (a.k.a. Best Linear Unbiased Estimator) can be used in order to find the optimal solution. Nevertheless, those estimators only works well under gaussian noise conditions that usually are not satisfied in real scenarios.

Table 1.1: Comparison between different methods

	Beamforming	TDOA	High-Resolution
Computational Complexity	High	Low	Medium
Robust against Reverberation	Yes	No	No
Multi-Source	Poor Results	No	Yes
Bandwidth	Wideband	Wideband	Narrowband†

† can be extended to wideband but increasing its computational cost

1.2.3 High Resolution Spectral Estimation

Methods based on spectral estimation of the signal, like the popular multiple signal classification algorithm (MUSIC) [29], exploit the spectral decomposition of the covariance matrix of the incoming signals for improving the spatial resolution of the algorithm in a multiple source context. These methods tend to be less robust than beamforming methods [2], and are very sensitive to small modeling errors.

Unlike SRP and its derivatives, incoherent signals are assumed by MUSIC, but in real scenarios with speech sources and reverberation effects, the incoherence condition is not fulfilled, making the subspace-based techniques problematic in practice.

Table 1.1 shows a summary of the previously described methods. This table is not exhaustive, and some methods in the literature combine the properties of each category.

1.3 Thesis Contributions

In this Thesis, we have made several contributions about acoustic source localization in reverberant environments. The contributions can be categorized into two big groups based in the previously described approaches.

The first category, which contains the majority of contributions, is related with beamforming techniques, GCC-PHAT and SRP-PHAT methods. On the other hand, the second category are the contributions to TDOA methods. The main difference between the two groups is that while all the techniques in the first group make use of all the samples in correlation (*i.e.* GCC-PHAT), the techniques based in TDOA discard most of this information, keeping only the TDOA estimations (*e.g.* peaks in correlation).

1.3.1 Based on PHAT

The main idea in this field is that source localization admits a sparse representation. Given that the space in a room is discretized in voxels, most of those voxels will thus be empty and only few of them will contain a source.

In order to exploit sparsity, we propose to use a linear generative model. This model expresses SRP-PHAT measurements (also can be used with GCC-PHAT directly) as the

linear combination of the contribution of each source. The position of the sources will be estimated as those that minimize the fitting error between SRP-PHAT measurements and the response predicted by the model. We augment this fitting process with a sparsity constraint whose aim is to impose that the number of sources is small on the search space.

Assuming that the position of the sources are constrained to a finite set $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_Q\}$ of Q positions, the former problem can be reformulated as:

$$\min_{\mathbf{a}} \|\phi - \Phi\,\mathbf{a}\|^2 \quad \text{s.t.} \quad \|\mathbf{a}\|_0 \leq k \ , \tag{1.1}$$

where k ($k \leq Q$) is the maximum number of speakers and $\Phi = [\bar{\phi}(\mathbf{q}_1), \cdots, \bar{\phi}(\mathbf{q}_Q)]$ is a matrix which contains the models for each of the possible locations.

Consequently, the support of the optimal $\mathbf{a}$, $\mathbf{a}^*$, is related to the estimated position of the sources. Thus, if $i \in \text{supp}(\mathbf{a}^*)$ then, our method estimates that a source is placed at $\mathbf{q}_i$.

Nevertheless this is a NP-hard and non-convex problem. Despite its theoretical complexity, several methods and approximations have been proposed. Especially relevant approximation is the used during this Thesis which is based on using the l_1 norm as a convex relaxation of the l_0 norm [10, 30]. This relaxation transforms equation (1.1) into the following:

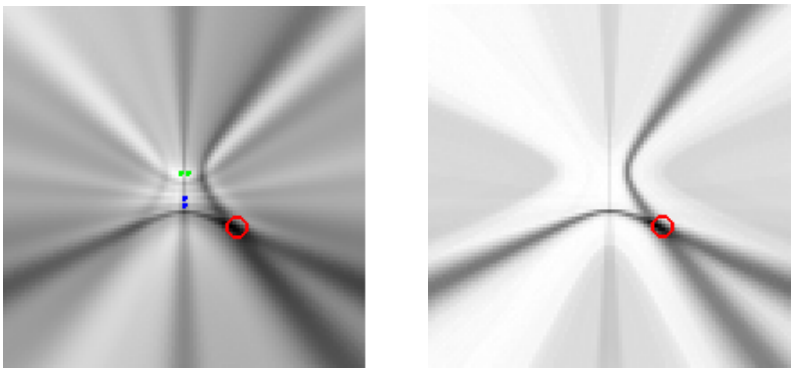
$$\min_{\mathbf{x}} \|\phi - \mathbf{\Phi a}\|^2 + \lambda \|\mathbf{a}\|_1 \tag{1.2}$$

where λ is the Lagrange multiplier and has a direct relationship with k. Problem (1.2) is convex, thus convergence is guaranteed and can be solved in polynomial time.

1.3.1.1 SRP-PHAT Denoising

Our first proposal [4] was a new approach for denoising SRP-PHAT based on using a generative model and sparsity constrains introduced above. Although that method does not take full advantage of the sparse representation (i.e. we were not able to localize directly from $\text{supp}(\mathbf{a}^*)$), we proved that l_1 constrained optimization was essential for performing denoising. Results showed statistically significant localization error reductions of up to 30% when compared to standard SRP-PHAT strategies, especially when only few pair of microphones are used.

Figure 1.1 shows the original SRP-PHAT power map and its denoised version which is calculated as $\mathbf{Phia}^*$. It seems clear that denoising effectively reduces the number of artifacts and unwanted effects exhibited by the original map. This yields a better detection of local maxima truly representing active acoustic sources.



(a) measured SRP-PHAT (b) Denoised SRP-PHAT

Figure 1.1: Comparison between real SRP-PHAT power map and its denoised version.

1.3.1.2 Modeling of the PHAT Filtering Effects

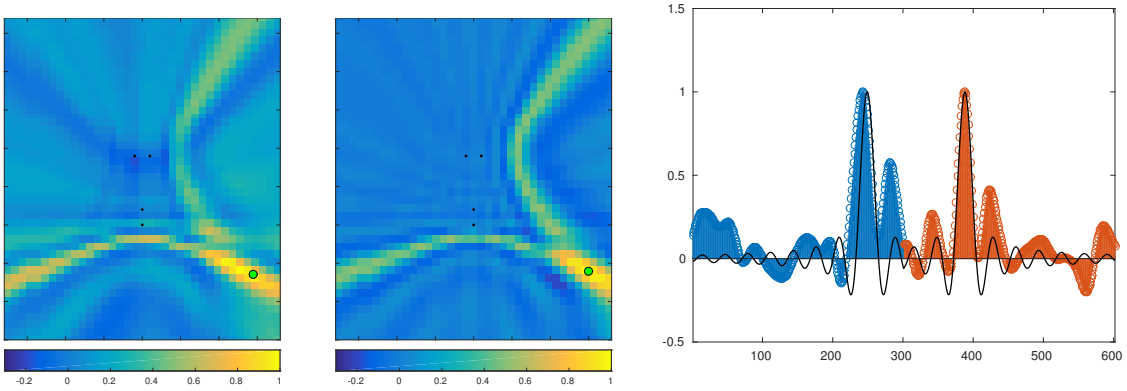
The model used in the previous approach [4] is overly coarse, not being able to represent accurately most of the SPR-PHAT artifacts. Therefore, such model was not enough to sparsely represent SRP-PHAT.

In this Thesis we describe and model the effects of PHAT filtering with low-band signals [14]. This yields an analytical model which allows us to predict with high accuracy the GCC-PHAT of the signals. Given the array geometry the model can be also extended to SRP-PHAT.

The proposed model is independent to the emitted signal (under some mild conditions) and has been shown to be valid in reverberant environments and under far and near field conditions (see Figure 1.2). Our model allows us to predict how the aforementioned factors affect the SRP-PHAT power maps. These predictions are validated with both synthetic and real data, showing that our model accurately reproduces SRP-PHAT power maps in both anechoic and non-anechoic scenarios. It is thus an excellent tool to be exploited for the improvement of real world relevant applications related to acoustic localization and calibration [15]. Furthermore, this model has been demonstrated to be useful predicting the optimal microphone placement for indoor acoustic localization [31].

1.3.1.3 Sparse Acoustic Source Localization

The last contribution of this Thesis shows that the proposed model in [14] is well suited for localization using sparse constraints. In this case we no longer work over SRP-PHAT representation but over GCC-PHAT. Working over GCC-PHAT allows us to reduce the computational cost by removing the need of computing SRP-PHAT from GCC-PHAT. Therefore, the input data is the concatenation of correlations (see Figure 1.2c) which is very convenient since the size of such vector is related with the number of microphones and the distance between them but not to the desired resolution.

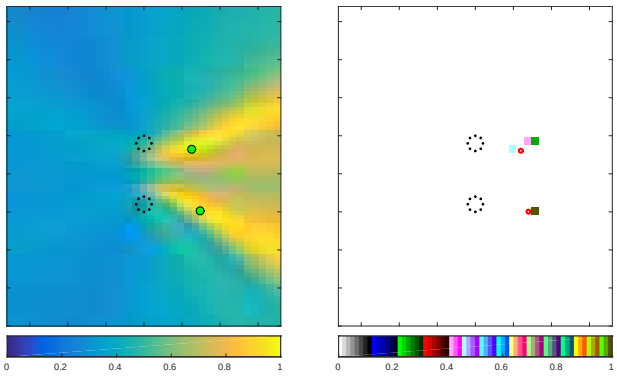


(a) SRP-PHAT (b) Pattern (c) Correlations

Figure 1.2: Comparison between the measured SRP-PHAT power map (Fig. a) and the SRP-PHAT power map predicted by the proposed model (Fig. b) using only two pair of microphones. Fig. c shows the measured GCC-PHAT in blue and orange (each color represents a different pair of microphones) and the corresponding model in black solid line.

In our preliminary results, where a single source is assumed, we outperform SRP-PHAT by using a very low-complexity algorithm that derives from our model and sparse constraints based on the l_0 regularization. We also expand the model proposed in [14] for multiple sources demonstrating its linear behaviour when the coherence between sources is small. Preliminary experiments shows promising results with multiple sources by finding a sparse representation based on the l_1 regularization.

In figure 1.3b is shown the solution of the problem (1.2) when two speakers are active. Sparse representation is more compact, and therefore convenient, than the classical SRP-PHAT power Map (figure 1.3a). In multiple source scenario more pairs of microphones are needed for an accurate localization. In both images microphones have been represented as black dots (shaped in circles in the center of the images).



(a) Measured SRP (b) Sparse solution

Figure 1.3: Comparation between SRP-PHAT power map and our proposal for two sources

1.3.1.4 GCC-PHAT Model for Calibration in Diffuse Noise

We have also extended the model of GCC-PHAT for the diffuse noise case [15]. Diffuse noise is the kind of noise present in a quiet room or car (i.e. only with the noise of fans, air conditioned, computers...). It can be roughly described as an acoustic field where the signals propagate in all directions with the same power and equal probability.

The proposed model only depends on the pairwise distance between microphones and the signal bandwidth, as shown in figure 1.4 . It's thus suitable for geometry array calibration. Furthermore, the proposed method for calibration is more accurate and performs faster than other calibration methods based in the coherence of the diffuse noise [32].

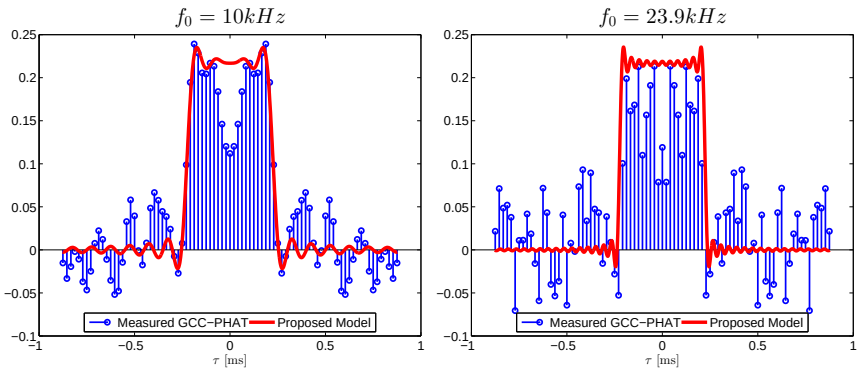


Figure 1.4: The proposed GCC-PHAT model compared with the measured GCC-PHAT on real data recordings in a diffuse sound field. The dependency on the signal bandwidth is demonstrated: the left graphic uses $f_0 = 10$ kHz and the right one uses $f_0 = 23.9$ kHz.

1.3.2 Based on TDOA: TDOA Matrices

Measuring TDOA between a set of sensors is the basic setup for many applications, such as localization or signal beamforming. TDOA Matrices [16] are a good representation for redundant TDOA measurements in sensor arrays. In this Thesis we have studied and given proofs of the properties of such matrices.

A TDOA matrix $\mathbf{M}$, is a $(n \times n)$ skew-symmetric matrix where the element (i, j) is the time difference of arrival (TDOA) between the signals arriving at sensor i and sensor j:

$$\mathbf{M} = \{\Delta\tau_{ij}\} = \begin{pmatrix} 0 & \Delta\tau_{12} & \cdots & \Delta\tau_{1n} \\ \Delta\tau_{21} & 0 & \cdots & \Delta\tau_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta\tau_{n1} & \Delta\tau_{n2} & \cdots & 0 \end{pmatrix} \tag{1.3}$$

with $\Delta\tau_{ij} = (\tau_i - \tau_j)$, where τ_i is the time of arrival of the signal $x(t)$ at the sensor $\mathbf{s}_i$.

We denote as $\mathcal{M}_T(n)$ to the set of TDOA matrices of size $n \times n$.

Note that in the former definition, knowing the sensor array geometry is not required so that they can also be used in calibration [33]. For a given geometry, all the feasible TDOA matrices (those that are consistent with that particular geometry) are a subset of $\mathcal{M}_T(n)$.

Additionally, given a particular TDOA matrix, there are infinite number of sensors geometries which match with it. Left side of figure 1.5 shows that, given a set of TOAs (τ_1,..., τ_n) compatible with the set of TDOA measurements, the microphones can be situated in any place along the circumference (sphere in the 3D case) with center in the source (dotted lines), preserving its correspondent TOA (and therefore, its TDOA). Right side of figure 1.5 shows that there are an infinite number of TOA sets that comply with a given set of TDOA measurements.

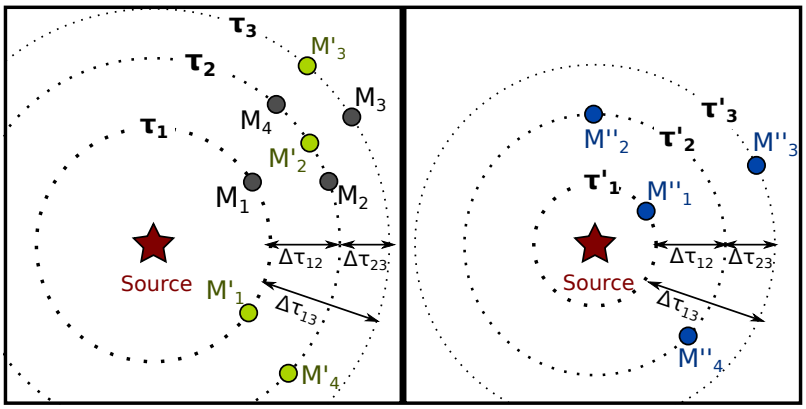


Figure 1.5: Example of three different geometrical configurations (grey, green and blue) of 4 sensor with identical TDOA matrix.

We demonstrate TDOA matrices are rank-two and have a special SVD decomposition that leads to a compact linear parametric representation. We apply these properties to perform denoising, by finding the TDOA matrix closest to the matrix composed with noisy measurements. The Thesis shows that this problem admits a closed-form solution for TDOA measurements contaminated with Gaussian noise which extends to the case of having missing data yielding the Gauss-markov estimator (a.k.a. Best linear unbiased estimator). We also propose a novel robust denoising method resistant to outliers, missing data and inspired in recent advances in robust low-rank estimation.

Several experiments with both synthetic and real data show significant improvements for the proposed denoising algorithms in terms of TDOA accuracy estimation and localization error. Furthermore, since knowledge about the sensor array geometry is not necessary for none of the proposed algorithms, they can also be used for calibration.

Chapter 2

Conclusions and Future Work

In this chapter we will summarize the most relevant conclusions derived from this thesis. Additionally, we also provide some proposal of future lines that could be tackled from this thesis.

2.1 Conclusions

During this thesis novel mathematical models have been developed for the problem of wideband acoustic source localization from microphone arrays. Unlike recent high resolution spectral models, such as MUSIC, we base our modelling on the GCC-PHAT and SRP-PHAT of the signals. These methods have demonstrated to have better performance in real scenarios.

We have demonstrated that measured GCC-PHAT, and consequently SRP-PHAT, can be accurately predicted by our analytical models wich has been proposed during this thesis. They include reverberation effects and far and near field conditions. Multisource scenario has also been addressed during this thesis. We have demonstrated that, if the coherence between sources is small, the contribution of each source in the measurements is linear.

Additionally, a new model for GCC-PHAT in diffuse sound field was proposed which establishes the links between GCC-PHAT output and the microphone array geometry. It was shown that this model is in fact equivalent to an inverse Fourier transform of an ideally filtered coherence of the two signals.

These models provide extra information that was unemployed until this thesis. Nevertheless, we propose several methods and algorithms to exploit it in many tasks. For instance, we exploit it to achieve higher resolution in localization or higher accuracy in calibrating microphone arrays from diffuse noise.

Concerning localization, we have explored the use of sparse constraints in source localization problems. This fits very well with the fact that generally only a few sources

are active at the same time. Sparse constraints are included in model-based fitting for SRP-PHAT and GCC-PHAT. Namely, we studied both l_0 constraints and their convex relaxations using l_1 regularization.

We have shown along this thesis that having an accurate model is necessary in order to get an sparse representation of the source localization problem. Nevertheless, a coarse model has been employed for SRP-PHAT denoising with very good results in localization.

We mainly focused on single source scenarios where we show that imposing sparsity improves localization accuracy. This thesis also presents some preliminary results in multisource scenarios, where we obtain promising results that require further theoretical analysis and are left as a future work.

Alternatively we have studied the algebraic properties of TDOA Matrices, an interesting representation for solving TDOA denoising problems. Using such properties we have addressed denoising of TDOA measurements contaminated with gaussian noise, outliers and even where a percentage of such measurements were missing. The experimental results, both on real and synthetic data have shown that our algorithms successfully perform denoising (up to 30% of improvement in localization accuracy) with a high rate of missing data (up to 50%) and outliers, without knowing the sensor positions. This is important as it can be applied to tasks where the sensors geometry is unknown. Interestingly, in real datasets our robust denoising algorithm is systematically better than the Gauss-Markov estimator even when there is no missing data. This is also an important result as it proves that the assumption of Gaussian noise does not hold in real cases, while our robust model is capable of automatically discard erroneous measurements.

Finally, he experiments conducted on real data recordings demonstrate the effectiveness of the proposed calibration approach for pairwise distance estimation. The proposed model suggests a simple denoising scheme for the coherence function via suppression of the GCC-PHAT activation at the time intervals which do not meet the physical constraints. The model was shown to perform significantly faster than the coherence- based counterpart and it is applicable for real time calibration setups.

2.2 Future Work

1. Although, preliminary results are promising, more extensive experimentation is needed in the multispeaker scenario. It is necessary to perform a systematic experiment varying the number the number of sources as well as the number of microphones and their position.

2. We have define a linear basis for acoustic source localization. Each vector within such basis correspond with the GCC-PHAT expected response in each position of a given grid. During this thesis we haven't done any assumption about the location

of the grid positions. Nevertheless, it seems that an optimal criteria to design the grid could be employed, for instance, attending to the restricted isometry property (RIP). It will improve the sparse approximation of the l_1 regularizated problem.

3. It is well known that reverberation contains information about the room geometry that can be exploit. On the other hand, according with the proposed model, the reverberation effects can be predicted. Therefore, analyzing the residual after localization may throw some information about the room geometry.

4. During this thesis we have used a convex relaxation of the sparse constrained problem. This, is very common in compressive sensing where the amount of acquired data is drastically reduced by randomization of the input. It would be interesting investigate if it is possible to reduce the amount of data necessary to localize randomizing is some way the calculation of GCC-PHAT.

5. It would be also interesting exploring new optimization techniques for improving the results of this thesis.

6. Finally, a distributed version of our algorithms would improve their the applicability. For example, some smartphones in a meeting could calibrate themselves and localize sources in order to apply other kind of techniques such as beamforming. Maybe, it could be done using message passing and/or stochastic gradient descend algorithms, typically employed in distributed systems.

Bibliography

[1] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

[2] J. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180.

[3] A. Waibel and R. Stiefelhagen, *Computers in the Human Interaction Loop*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[4] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, pp. 13 781–13 812, 10/2012 2012.

[5] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, 1999.

[6] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[7] V. Temlyakov, "Nonlinear methods of approximation," *Foundations of Computational Mathematics*, vol. 3, no. 1, pp. 33–107, 2003.

[8] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[9] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, pp. 129–159, 2001.

[10] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[11] J. Claerbout and F. Muir, "Robust modeling with erratic data," *Geophysics*, vol. 38, p. 826, 1973.

[12] R. Baraniuk, "Compressive sensing [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 118–121, 2007.

[13] E. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[14] J. Velasco, C. J. Martín-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Processing*, vol. 119, pp. 209 – 228, 2016.

[15] J. Velasco, M. J. Taghizadeh, A. Asaei, H. Bourlard, C. J. Martín-Arguedas, J. Macias-Guarasa, and D. Pizarro, "Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2669–2673.

[16] J. Velasco, D. Pizarro, J. Macias-Guarasa, and A. Asaei, "TDOA matrices: Algebraic properties and their application to robust denoising with missing data," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5242–5254, Oct 2016.

[17] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Procceedings of the MLMI*, ser. Lecture Notes in Computer Science, S. Bengio and H. Bourlard, Eds., vol. 3361. Springer-Verlag, 2004, pp. 182–195.

[18] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication*, ser. Springer Topics in Signal Processing, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer Berlin Heidelberg, 2010, vol. 3, pp. 307–337.

[19] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[20] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 288–292, 1993.

[21] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510 –2526, nov. 2007.

[22] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, oct. 2009, pp. 2033 –2038.

[23] H. Do and H. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proceedings of ICASSP 2010*, march 2010, pp. 125 –128.

[24] M. Cobos, A. Marti, and J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 1, pp. 71–74, 2011.

[25] T. Butko, F. Gonzalez Pla, C. Segura Perales, C. Nadeu CamprubÃ, and F. J. Hernando PericÃ¡s, "Two-source acoustic event detection and localization: online implementation in a smart-room," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO'11)*, 2011, pp. 1317–1321.

[26] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320 – 327, aug 1976.

[27] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proceedings of ICASSP 2008*, 31 2008-april 4 2008, pp. 2565 –2568.

[28] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[29] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276 – 280, mar 1986.

[30] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589 – 602, 2006.

[31] R. Macho-Pedroso, F. Domingo-Perez, J. Velasco, C. Losada-Gutierrez, and J. Macias-Guarasa, "Optimal microphone placement for indoor acoustic localization using evolutionary optimization," in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Oct 2016, pp. 1–8.

[32] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 3, pp. 666–670, 2008.

[33] Y. Kuang and K. Astrom, "Stratified sensor network self-calibration from TDOA measurements," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[34] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, pp. 13 781–13 812, 10/2012 2012.

Part II

Dissertation by compendium of publications: Articles

This second part of the manuscript presents the three articles as they have been published in indexed journals, the paper presented in a prestigious conference, and a technical report on the latest Thesis work, currently being used to prepare another journal publication. Each article/report address one of the formerly described topics:

Chapter 3: Journal paper on SRP Denoising

Chapter 4: Journal paper on Modeling of the PHAT Filtering Effects

Chapter 5: Technical Report on Sparse Acoustic Source localization (to be published)

Chapter 6: Conference paper on GCC-PHAT Model for Calibration in Diffuse Noise

Chapter 7: Journal paper on TDOA Matrices

Chapter 3

Journal Paper on SRP-PHAT Denoising: *Source Localization with Acoustic Sensor Arrays Using Generative Model Based Fitting with Sparse Constraints*

Publication reference:

- J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, pp. 13 781–13 812, 10/2012 2012

Sensors **2012**, *12*, 13781-13812; doi:10.3390/s121013781

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Source Localization with Acoustic Sensor Arrays Using Generative Model Based Fitting with Sparse Constraints

Jose Velasco *, Daniel Pizarro and Javier Macias-Guarasa

Department of Electronics, University of Alcalá, Campus Universitario s/n, 28805, Alcalá de Henares, Madrid, Spain; E-Mails: pizarro@depeca.uah.es (D.P.); macias@depeca.uah.es (J.M.-G.)

* Author to whom correspondence should be addressed; E-Mail: jose.velasco@depeca.uah.es; Tel.: +34-918-856-918, Fax: +34-918-856-591.

Received: 30 July 2012; in revised form: 25 September 2012 / Accepted: 26 September 2012 / Published: 15 October 2012

Abstract: This paper presents a novel approach for indoor acoustic source localization using sensor arrays. The proposed solution starts by defining a generative model, designed to explain the acoustic power maps obtained by Steered Response Power (*SRP*) strategies. An optimization approach is then proposed to fit the model to real input *SRP* data and estimate the position of the acoustic source. Adequately fitting the model to real *SRP* data, where noise and other unmodelled effects distort the ideal signal, is the core contribution of the paper. Two basic strategies in the optimization are proposed. First, sparse constraints in the parameters of the model are included, enforcing the number of simultaneous active sources to be limited. Second, subspace analysis is used to filter out portions of the input signal that cannot be explained by the model. Experimental results on a realistic speech database show statistically significant localization error reductions of up to 30% when compared with the *SRP-PHAT* strategies.

Keywords: acoustic localization; microphone array sensors; sparse modeling; optimization techniques

1. Introduction

The development and scientific research in perceptual systems has notably grown during the last decades. The aim of perceptual systems is to automatically analyze complex and rich information taken

from different sensors. These systems stem from basic sensor technologies, reaching the knowledge frontier in signal processing and pattern recognition research areas.

On top of perceptual systems, the idea of using sensors to analyze the real world has emerged in different scientific disciplines such as "ubiquitous computing" [1], "smart rooms" [2] or "intelligent spaces" [3]. All these disciplines lay stress on the idea of systems with interaction capabilities that can analyze human activities and provide services.

A basic but important milestone inside these disciplines is the development of sensor technologies able to localize humans in indoor environments. Localization of humans has a tremendous potential impact in diverse applied fields, opening new ways in how humans interact with machines. One important factor in indoor localization is the user awareness of the sensors used. Non-invasive technologies are preferred in this context, so that no electronic or passive devices are to be carried by humans for localization. The two non-invasive technologies that have been mainly used in indoor localization are those based on video systems and acoustic sensors.

Video systems provide very rich information at a low cost on the sensor side. However, video analysis is a complex problem and needs a lot of effort to build robust and reliable systems. In recent years, there are many publications focused on video-based indoor localization systems for humans [4,5], robots [6], and object recognition systems [7].

Acoustic sensors give also very rich information as humans communicate mainly with speech. As in video, there is also a considerable amount of publications focused on obtaining the exact position of any active acoustic source in a scene [8,9]. Video and audio technologies are in fact very complementary in many ways [10].

This paper focuses on audio-based localization in a very general scenario, where unknown wide-band audio sources (e.g., human voice) are captured by a set of microphone arrays placed in known positions. The main objective of the paper is to use the signals captured by the microphone arrays to automatically obtain the position of the acoustic sources detected. Especially relevant in practice are the methods based on computing the Steered Response Power (*SRP*) [11] of the signals captured in microphones arrays. These approaches have proved to be successful for localization in reverberant and noisy scenarios [12].

This paper proposes a simple generative model to explain *SRP* measurements in environments equipped with any combination of microphone arrays. The main contribution of the paper is to use an optimization approach to fit the generative model to noisy *SRP* data, exploiting the fact that only a few speakers are expected to be active at the same time. This simple idea is modeled with sparse constraints in the optimization cost, and combined with subspace filtering. The paper shows that this model-based approach can be used to notably improve the localization results of the state-of-the-art methods based on *SRP-PHAT*. Although this proposal is developed and evaluated for speech signals, the authors believe that it is general enough to be easily extended to other wideband and narrowband acoustic signals.

1.1. Paper Structure

The paper is structured as follows. In Section 2 we provide an extensive study of the state-of-the-art in acoustic source localization and optimization methods. Section 3 describes the proposed generative model and Section 4 deals with the optimization strategy to fit the model to real data. The experimental

evaluation is detailed in Section 5, and Section 6 summarizes the main conclusions and contributions of the paper and gives some ideas for future work.

2. State of the Art

2.1. Acoustic Source Localization

The acoustic source localization methods are the starting point of other techniques like speech enhancement using beamforming. Therefore, acoustic source localization has received significant attention lately as a mode of automatic tracking of persons and as a complement to other existing alternatives of tracking, e.g., the CHIL (Computer in Human Interaction Loop) project [10].

Many approaches exist in literature and all of them use microphone arrays as a non-intrusive method. These can roughly be divided in three categories [8,9]: time delay based, beamforming based, and high-resolution spectral-estimation based methods.

The first methods are based on estimating the time delay of signals relative to pairs of spatially separated microphones. Assuming uncorrelated, stationary Gaussian signal and noise with known statistics and not multi-path, the maximum likelihood (ML) time-delay estimate is derived from a SNR-weighted version of the Generalized Cross Correlation (GCC) function [13]. In a second step, the time-difference of arrival information is combined with knowledge of the microphones' positions to generate a ML spatial estimator made from hyperbolas intersected in some optimal sense [8,9].

An accurate estimation of the time delay is essential for a good performance of this *time delay of arrival* (TDOA) methods. Since coherent noise and multi-path due to reverberation are the two major sources of error in time delay estimation, different approaches have been proposed to deal with them. A basic method consists in making the GCC function more robust, de-emphasizing the frequency-dependent weighting. The Phase Transform (PHAT) [13] is one example of this procedure that has received considerable attention as the basis of speech source localization systems due to its robustness in real world scenarios [14].

Beamforming based techniques [15] attempt to estimate the position of the source, maximizing or minimizing a spatial statistic associated with each position. For instance, in the Steered Response Power (*SRP*) approach, which is the simplest beamforming method, the statistic is based on the signal power received when the microphone array is steered in the direction of a specific location. Therefore, the position of the source is supposed to be consistent with the position corresponding to the maximum estimated signal power

SRP-PHAT is a widely used algorithm for speaker localization based on beamforming. It was first proposed in [11] and is a beamforming based method that combines the robustness of the steered beamforming methods with the insensitivity to signal conditions afforded by the Phase Transform (PHAT). The classical delay-and-sum beamformer used in *SRP* is replaced in *SRP-PHAT* by a filter-and-sum beamformer using PHAT filtering to weight the incoming signals. In this paper, the term *SRP* will be used interchangeably with *SRP-PHAT*.

The advantage of using PHAT is that no assumptions are made about the signal or room conditions [16], and this is the reason for the robustness of the *SRP-PHAT* method in reverberant scenarios, where the source is unknown. *SRP-PHAT* is usually defined as a reference standard for

source localization, because of its simplicity and robustness in reverberant and noisy environments, being a widely used algorithm for speaker localization [17–21].

The Minimum Variance Distortionless Response (MVDR), also called Capon's method, is another beamforming based approach which takes advantage of the estimated signal and noise parameters. These parameters are used to carry out optimal beamforming techniques in order to minimize the measured power from noise and sources located in other positions. However, MVDR has a poor performance in the presence of reverberation, because it introduces a new trade-off between de-reverberation and noise reduction [22] .

In [23,24], a unified maximum likelihood framework is presented, which is equivalent to forming multiple MVDR beamformers along multiple hypothesis directions and picking the output direction which results in the highest SNR [24]. Apparently, it outperforms *SRP-PHAT* in reverberant real scenarios.

The spectral estimation based methods, like the popular multiple signal classification algorithm (MUSIC) [25], exploit the spectral decomposition of the covariance matrix of the incoming signals for improving the spatial resolution of the algorithm in a multiple sources context. These methods tend to be less robust than beamforming methods [9], and are very sensitive to small modeling errors.

Unlike *SRP* and its derivatives, incoherent signals are assumed by MUSIC, but in real scenarios with speech sources and reverberation effects, the incoherence condition is not fulfilled, making the subspace-based techniques problematic in practice.

The work presented in this paper uses *SRP-PHAT* as the base to develop a generative model to explain real data, and the experimental results are compared against *SRP-PHAT*.

2.2. Sparse Representation of Signals

Many areas of science share the principle of parsimony as the central criterion: the simplest explanation of a given phenomenon is preferred over more complicated ones. This brilliant idea has been recently applied to the representation of signals using overcomplete basis sets, sometimes called dictionaries in the machine learning discipline. As a difference with respect to traditional basis functions (e.g., Fourier basis functions), overcomplete dictionaries have more degrees of freedom than those necessary to represent the signal. The mathematical tool to impose parsimony in the representation of a signal, when several choices are available, is given by imposing the so-called sparse constraints. The basic idea is to use the least amount of coefficients to represent a signal with the basis functions. Sparse constraints, if they are applicable, allow to beat up several theoretical barriers in signal compression and representation [26,27].

The sparsity is imposed mainly by using optimization approaches, where the l_0 norm (defined as the number of non-zero elements in the vector) is the usual way to impose sparsity to vectors [27].

Most of the problems in which sparsity is included using the l_0 norm are very difficult to solve. Several methods have been proposed to find sparse representations, including brute force approaches as well as more computationally efficient approximate methods such as "nonlinear programming" [28], and greedy pursuit [29–31]. Among all approximate solutions, l_1 norm based convex relaxations have flourished in the literature. The Basis Pursuit method [32,33], originally introduced by [34] almost 40 years ago but

revisited with a profound theoretical study in the past decade, can be highlighted due to its intensive use in the modern compressive sensing techniques [26,27]. These methods provide very effective polynomial time algorithms that, under certain circumstances, are even equivalent to the original l_0 based problems [27,33].

2.3. Sparse Source Localization

In the last few years, sparse techniques explained above have been applied to the source localization problem in very different fashions.

In [35] a localization approach based on sensor arrays is proposed. The signal obtained in each sensor is expressed as a linear combination of an attenuated and phase shifted version of the original and known signals emitted by the source. This conditions form an overcomplete linear model, where the position of the sources is given thanks to the sparse constraints. Also in [35] they propose to use *singular value decomposition* (SVD) to reduce problem size and filter noise in problems using multiple time samples.

The work presented in this paper includes sparse and SVD decompositions for acoustic source localization but the objectives (unknown source signals) and the way these techniques are applied are very different to those of [35]. Our proposal works in the *SRP-PHAT* acoustic power maps, while [35] operates at the sensor signal level.

Numerous modifications of the ideas proposed in [35] has been further developed. For example, in [36] an adaptive algorithm to dynamically adjust both the overcomplete basis and the sparse solution is proposed. Also, the concept of Compressive Sensing [27] has been used in order to perform a distributed localization reducing the information transmitted between sensors. Nevertheless, the sparse source localization algorithms discussed above do not perform well and are not properly tested in real acoustic reverberant environments due to input signals coherence caused by multipath.

In acoustic environments, sparse l_1 relaxations are employed to model the room acoustically using only a reduced number of microphones in [37]. However, only simple rooms (four walls and ceiling) can be modeled, and a loudspeaker emitting a known sound pattern is required. Using this technique in a previous training step has been proved to be useful to improve source localization [38].

Recently, a novel technique for source localization in reverberant environments using wavefield sparse decomposition has been proposed in [39]. However, although it shows promising performance, the experimental results are only based on simulations and narrowband signals, which makes their approach not applicable to speech signals, which is our target scenario.

3. Model Proposal

3.1. Notation

Real scalar values are represented by lowercase letters (e.g., δ). Upper-case letters are reserved to define vector and set sizes (e.g., vector $\mathbf{x} = (x_1, \cdots, x_N)^\top$ is of size N). Vectors are by default arranged column-wise and are represented by lowercase bold letters (e.g., $\mathbf{x}$). Matrices are represented by uppercase bold letters (e.g., $\mathbf{M}$). The l_p norm ($p > 0$) of a vector is depicted as $\|.\|_p$, e.g., $\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_N|^p)^{\frac{1}{p}}$, where $|.|$ is reserved to represent absolute values of scalars.

Special cases are the l_0 norm, written $\|.\|_0$ and defined as the number of non-zero elements in the vector, and the l_∞ norm, written $\|.\|_\infty$ and defined as the maximum value of the vector components. The l_2 norm $\|.\|_2$ will be written by default as $\|.\|$ for simplicity. Calligraphic fonts are reserved to represent sets (e.g., $\mathbb{R}$ for real or generic sets $\mathcal{G}$).

3.2. Interpretation of the SRP-PHAT *Estimations*

Assume we have equipped a certain indoor environment with a set of N different microphone pairs distributed in some fashion in three-dimensional known positions. All pairs of microphones are described as elements in a set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, where $\mathbf{p}_j = (\mathbf{m}_j, \mathbf{m'}_j)$ is composed of two three-dimensional vectors, $\mathbf{m}_j$ and $\mathbf{m'}_j$, describing the spatial location of the microphones in pair j.

The three-dimensional space where acoustic sources are to be localized is discretized using a finite set of Q spatial locations $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_Q\}$, where $\mathbf{q}_k$ is a three-dimensional vector $\mathbf{q}_k = (q_{kx}, q_{ky}, q_{kz})^\top$.

The classical *SRP-PHAT* method constructs a statistic $srp(\mathbf{q}_k), \mathbf{q}_k \in \mathcal{Q}$ based on the steered power received by all pairs of microphones from each spatial location. Simplifying the mathematical description of the *SRP-PHAT* formulation of [11] and applying the summation over all microphone pairs, we can write

$$srp(\mathbf{q}_k) = 2\pi \sum_{\forall \mathbf{p}_j \in \mathcal{P}} c_j(\Delta\tau(\mathbf{p}_j, \mathbf{q}_k)) \tag{1}$$

where $c_j(\Delta\tau(\mathbf{p}_j, \mathbf{q}_k))$ is the generalized cross-correlation (generally applying a PHAT weighting) of the signals acquired by each microphone in the pair $\mathbf{p}_j$, and

$$\Delta\tau(\mathbf{p}_j, \mathbf{q}_k) = \frac{1}{c}\left(\|\mathbf{m}_j - \mathbf{q}_k\| - \|\mathbf{m'}_j - \mathbf{q}_k\|\right) \tag{2}$$

is the difference in arrival times of the audio signal to reach microphones $\mathbf{m}_j$ and $\mathbf{m'}_j$, that is, the required delay to steer the microphone pair $\mathbf{p}_j$ to the location $\mathbf{q}_k$. In Equation (2) c is the sound velocity in air. Note than in the *SRP-PHAT* formulation we do not make any assumption regarding near-field/far-field conditions.

So, Equation (1) shows how the *SRP-PHAT* power estimation for every location $srp(\mathbf{q}_k)$ can be calculated as the sum of the cross-correlation functions for all microphone pairs, evaluated at the adequate steering delays (full implementation details of *SRP-PHAT* can be found in [11]). It is thus expected to see high values of $srp(\mathbf{q}_k)$ in regions in which active acoustic sources exist.

To provide an easier geometric interpretation, we now restrict the result of the $srp(\mathbf{q_k})$ estimations when only one omnidirectional acoustic source is active at position $\mathbf{s} = (s_x, s_y, s_z)^\top$, and only one microphone pair, e.g., pair $\mathbf{p}_j$, is located in the environment. The *SRP-PHAT* power estimation at $\mathbf{s}$ can be calculated as:

$$srp(\mathbf{s}) = 2\pi c_j(\Delta\tau(\mathbf{p}_j, \mathbf{s})) \tag{3}$$

From Equation (3), if we define $\mathbf{q}_h$ as the locations in $\mathcal{Q}$ for which $\Delta\tau(\mathbf{p}_j, \mathbf{q}_h) = \Delta\tau(\mathbf{p}_j, \mathbf{s})$, the corresponding cross-correlation values $c_j(\Delta\tau(\mathbf{p}_j, \mathbf{q}_h))$ will be identical to $c_j(\Delta\tau(\mathbf{p}_j, \mathbf{s}))$, consequently:

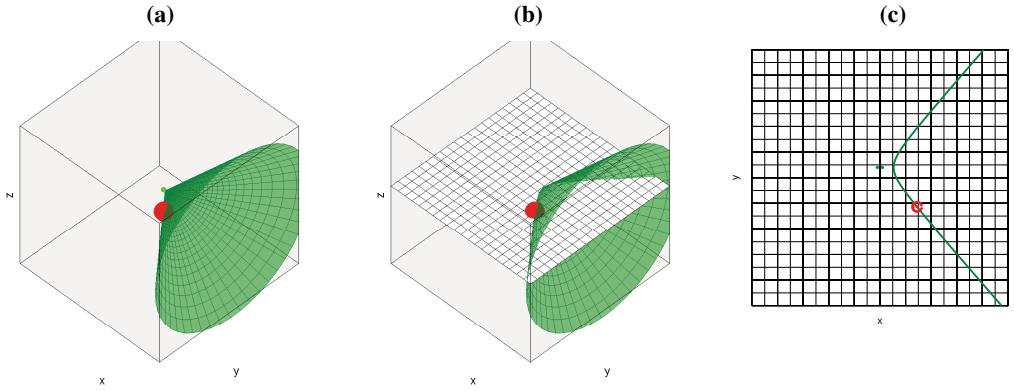
$$srp(\mathbf{q}_h) = srp(\mathbf{s}) \quad \text{if} \quad \Delta\tau(\mathbf{p}_j, \mathbf{q}_h) = \Delta\tau(\mathbf{p}_j, \mathbf{s}) \tag{4}$$

For a microphone pair, it can be easily demonstrated that the geometric place of points $\mathbf{q}_h$, for which the difference in time delays of arrival to the position of two microphones ($\Delta\tau(\mathbf{p}_j, \mathbf{q}_h)$ in our case) is equal to a given fixed value ($\Delta\tau(\mathbf{p}_j, \mathbf{s})$ in our case), is one of the sheets of a two-sheeted hyperboloid of revolution, whose foci are located at the microphone locations, as shown in Figure 1(a). If we define $\mathcal{H}$ as all the points $\mathbf{q}_h$ in $\mathcal{Q}$ that *belong* to the hyperboloid that *passes* through the acoustic source location $\mathbf{s}$, the *ideal SRP-PHAT* power estimation for all points in $\mathcal{Q}$ will be:

$$srp(\mathbf{q}_k) = \begin{cases} srp(\mathbf{s}) & \forall \mathbf{q}_k \in \mathcal{H} \\ 0 & otherwise \end{cases} \tag{5}$$

Equation (5) is correct if we assume that the environment is not reverberant and the array directivity pattern is perfect (*i.e.*, maximum gain in the steered direction and perfect cancellation in all other directions). We will address the effect of these simplifications in Section 3.3.

Figure 1. Geometric places with equal $srp(\mathbf{q}_h)$ generated for a microphone pair and a single acoustic source (a) 3D hyperboloid; (b) 3D hyperboloid cut by a plane; (c) Resulting 2D hyperbola (cutting hyperboloid by a plane).



Further simplifying, if we restrict the $\mathbf{q}_k$ positions to be located in a plane at a given height in the environment ($q_{kz} = z_0 \quad \forall \mathbf{q}_k \in \mathcal{Q}$), then $srp(\mathbf{q}_k)$ can be easily represented as an image that can be interpreted as the scene *acoustic power map*. In this situation, the place of points $\mathbf{q}_k$ with power equal to $srp(\mathbf{s})$ will be the result of *intersecting* the proper sheet of the hyperboloid of revolution with a plane parallel to the environment floor at z_0, and the generated geometric figure obtained will be a hyperbola.

As an example, if we consider the case of microphone pair $\mathbf{p}_j$, composed of microphones $\mathbf{m}_j = (-f, 0, 0)$ and $\mathbf{m'}_j = (f, 0, 0)$, and given a time difference of arrival $\Delta\tau(\mathbf{p}_j, \mathbf{s}) = \frac{1}{c}\left(\|\mathbf{m}_j - \mathbf{s}\| - \|\mathbf{m'}_j - \mathbf{s}\|\right)$ for a speaker position $\mathbf{s}$, the feasible acoustic source locations $\mathbf{q}_h = (x, y, z) \in \mathcal{Q}$ are those which satisfy the following expression (from Equations (2)–(4)):

$$\Delta\tau(\mathbf{p}_j, \mathbf{q}_h) = \frac{1}{c}\left(\|\mathbf{m}_j - \mathbf{q}_h\| - \|\mathbf{m'}_j - \mathbf{q}_h\|\right) = \Delta\tau(\mathbf{p}_j, \mathbf{s}) \tag{6}$$

Condition (6) defines the place of feasible locations $\mathbf{q_h}$ to be located in one sheet of the following two-sheeted hyperboloid of revolution (shown in Figure 1(a)):

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{b^2} = 1 \tag{7}$$

where a and b are related to the corresponding time difference of arrival $\Delta\tau(\mathbf{p}_j, \mathbf{s})$ and the microphones position through the following expressions:

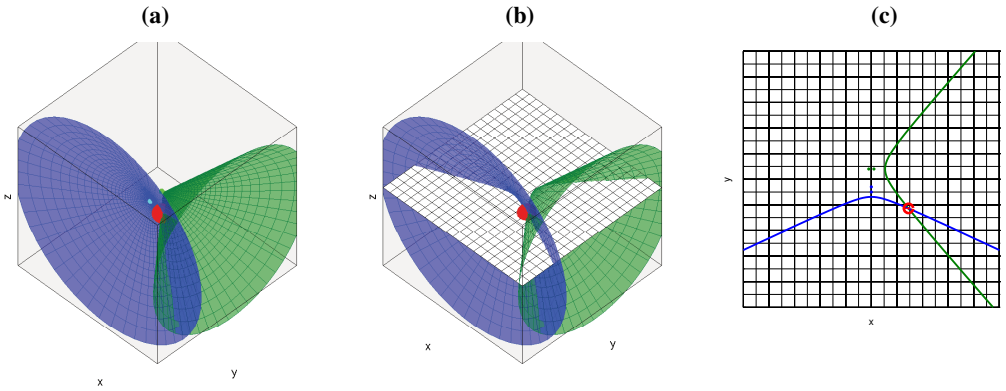
$$a = c\Delta\tau(\mathbf{p}_j, \mathbf{s})/2 \tag{8a}$$

$$b^2 = f^2 - a^2 \tag{8b}$$

Figure 1(c) shows the hyperbola that results from intersecting the hyperboloid with a plane, as shown in Figure 1(b).

If we add additional microphone pairs, each of them will generate a new hyperboloid/hyperbola, all passing through the geometric location of the active acoustic source, as shown in Figure 2(a) for the 3D case and Figure 2(c) for the 2D case (cutting the hyperboloids by a plane as shown in Figure 2(b)). Using additional microphone pairs will allow us to disambiguate the actual position of the acoustic source, searching in the intersection of all hyperboloids/hyperbolas.

Figure 2. Geometric places generated for two microphone pairs and a single acoustic source (a) 3D hyperboloids; (b) 3D hyperboloids cut by a plane; (c) Resulting 2D hyperbolas (cutting hyperboloids by a plane).



The final conclusion of this section is that, given some simplifications, for every active acoustic source and every microphone pair, we will see hyperbolic regions of *constant* acoustic power values in the acoustic power map generated by the ideal *SRP-PHAT* estimations. All the contributions for every acoustic source and every microphone pair will sum up to build the complete acoustic power map for the given situation.

3.3. Considerations in Real-World Scenarios

The simplifications established in this discussion (namely, only one omnidirectional acoustic active source, an ideal directivity pattern for the acoustic sensor array, and a non-reverberant environment) are far from being admissible in a real world scenario and deserve an additional comment:

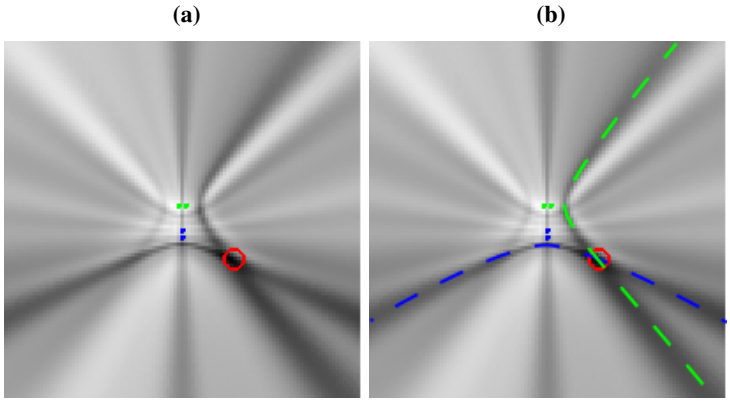
- Non-omnidirectionality of the acoustic active source: Previous studies such as [40] and [41] show that human speakers do not radiate speech uniformly in all directions. The impact of this assumption in our *SRP-PHAT* interpretation would lead to hyperbolic regions with different power estimations, but this effect is also present in the current formulation, as the distance between the acoustic source and the microphone varies with the source position. The use of the PHAT transform that *whitens* the correlation of the input signals alleviates this problem, as the module is not taken into account.

- Reverberant environments: If the localization system operates in a reverberant environment, new hyperbolic regions, not initially predicted by just the position of the acoustic source, will appear. Room acoustic simulation techniques could help in improving the ability to also take into account these regions [42,43]. These *false* active regions actually complicate the accurate location estimation, but the problem is alleviated as more microphone pairs are taken into account: locations that are not *consistent* for all microphone pairs will tend to attenuate. As we will see in Section 5, our proposal is actually efficient in *denoising* the original *SRP-PHAT* power map, thus leading to better results.

- Non-ideal directivity patterns: The microphone array geometry has a profound impact in the estimation of the cross-correlation functions, as the steered response will perceive energy coming from locations different from the actual acoustic source [44]. This implies that the acoustic power map will not be composed of plain hyperboloids/hyperbolas, but of hyperbolic *regions* spreading from the ideal hyperbolic trajectories, as will be shown in Figure 3(a), described in the next section. There are additional considerations that contribute to this *spreading* effect, related to the fact that the spatial uncertainty in the correlation evaluation increases as we move further from the microphone pairs. This will be addressed also in the next section.

To give a real world example, Figure 3(a) shows a real *SRP-PHAT* image generated by two microphone pairs (blue and green dots in the center of the image) and a single active speaker located at the red circle (the higher the power, the darker the color in the map). Analyzing this image, we can clearly see two high energy, intersecting hyperbolic areas passing trough the speaker location, each one corresponding to each microphone pair. Obviously, the speaker's position corresponds to the place where those hyperbolic areas intersect, as the maximum of the power map is found at this intersection. In general, the higher the number of microphone pairs used, the better the localization performance, as more hyperbolic regions contribute to the power map estimation. In Figure 3(b) the ideal hyperbolas corresponding to each of the microphone pairs have been superimposed to the *SRP-PHAT* map. The power map has been calculated at a plane located 61 cm above the microphone locations, which is why the hyperbolas do not *pass* between the hyperbola's foci—the microphone locations.

Figure 3. Real *SRP-PHAT* power map generated for a single speaker located in the red circle with two microphone pairs (blue and green dots). (a) Plain power map; (b) Superimposing ideal hyperbolas that should be generated by the single speaker.



This example shows us that in real acoustic power maps, the ideal hyperbolic functions are spread out and blurred, leading to these hyperbolic *areas*, and that additional hyperbolic areas appear, not explainable by just the position of the active acoustic source.

Summarizing, all these non-idealities will generate additional artifacts, additional hyperbolic regions and variations on the standard behavior of these regions in the acoustic power map that are not predicted by the ideal formulation. These non-idealities should be taken into account if we want our model to be as precise as possible. Our thesis is that our proposal, even when no developing a fully realistic model, is powerful enough to extract relevant information given realistic data, as will be shown in Section 5.

3.4. Proposal of a SRP-PHAT *Based Generative Model*

Taking into account the previous discussion and results, this section proposes a generative model that is able to explain the acoustic power map generated by *SRP-PHAT* as a sum of basis functions.

Let us define the set of scalar functions $\mathcal{F} = \{f(\mathbf{s}_i, \mathbf{p}_j, \mathbf{q}_k)\}, \quad \forall \mathbf{s}_i \in \mathcal{Q}, \quad \forall \mathbf{p}_j \in \mathcal{P}$, with $f : \mathbb{R}^{3\times6\times3} \rightarrow \mathbb{R}$. From this, the general formulation of the proposal can be written as:

$$\hat{srp}(\mathbf{q}_k) = \sum_{\forall \mathbf{s}_i \in \mathcal{Q}} \omega(\mathbf{s}_i) \sum_{\forall \mathbf{p}_j \in \mathcal{P}} f(\mathbf{s}_i, \mathbf{p}_j, \mathbf{q}_k) \tag{9}$$

where $\hat{srp}(\mathbf{q}_k)$ is the model estimation of $srp(\mathbf{q}_k)$, and the weights $\omega(\mathbf{s}_i)$ will be non-zero if there is an acoustic source in the given position $\mathbf{s}_i$, or 0 if otherwise.

The basis functions $f(\mathbf{s}_i, \mathbf{p}_j, \mathbf{q}_k)$ must be designed so that they provide accurate estimations of the behavior of the real *SRP-PHAT* value at location $\mathbf{q}_k$, taking into account that there is an active source at position $\mathbf{s}_i$ and that the signal is acquired by the microphone pair $\mathbf{p}_j$. This generic formulation allows for models (basis functions) as complex as required, in principle able to include any of the considerations described in Section 3.3.

In the experimental work described in Section 5, we are using a relatively simple model that is able to clearly outperform standard *SRP-PHAT* results. In our experiments, the basis functions $f(\mathbf{s}_i, \mathbf{p}_j, \mathbf{q}_k)$

describe if point $\mathbf{q}_k$ belongs to the hyperbolic region generated by an acoustic source $\mathbf{s}_i$ and a given pair of microphones $\mathbf{p}_j$:

$$f(\mathbf{s}_i, \mathbf{p}_j, \mathbf{q}_k) = \begin{cases} 1 & if \quad |\Delta\tau(\mathbf{p}_j, \mathbf{s}_i) - \Delta\tau(\mathbf{p}_j, \mathbf{q}_k)| \leq \epsilon \quad \epsilon \geq 0 \\ 0 & otherwise \end{cases} \tag{10}$$

where threshold ϵ accounts for the fact that in real-world scenarios there are uncertainties in measuring time delays as discussed in Section 3.3. Using $\epsilon > 0$, the width of the hyperbolic region is not constant, modeling the effect that can be clearly seen in Figure 3(a). In fact, the width increases with distance to the microphone pair, partly because for a given uncertainty (error) in the time delay estimation (due to the fact that we are using sampled signals), the spatial uncertainty (error in precisely assigning a correlation value to a given spacial location) increases as we consider positions further away from the microphone pair generating the hyperbolic region.

The model described by Equations (9) and (10) is valid to reproduce *SRP-PHAT* measurements, as the hyperbolic regions of the power maps are related to the high values of the Generalized Cross Correlation function of each pair of microphones [9]. Consequently the position of the hyperbolic regions is consistent with the time difference of arrival for each microphone pair given a certain speaker position.

3.5. Description of a Linear Model of SRP-PHAT

Using the model previously proposed in Equation (9) over all positions inside $\mathcal{Q}$ the following vector $\hat{\mathbf{y}}$ is defined:

$$\hat{\mathbf{y}} = \begin{pmatrix} s\hat{r}p(\mathbf{q}_1) & \cdots & s\hat{r}p(\mathbf{q}_Q) \end{pmatrix}^\top \qquad \mathbf{q}_k \in \mathcal{Q} \tag{11}$$

This section shows that vector $\hat{\mathbf{y}}$ can be represented as a linear combination of vectors of size Q. Each vector is only representative of a specific spatial location where an acoustic source can be active. As was described in previous sections, this model accounts for the fact that single acoustic sources are viewed in *SRP-PHAT* data as the intersection of multiple hyperbolic regions.

For each position $\mathbf{q} \in \mathcal{Q}$, define the following vector $\mathbf{v}(\mathbf{s})$:

$$\mathbf{v}(\mathbf{s}) = \begin{pmatrix} v(\mathbf{s}, \mathbf{q}_1), \cdots, v(\mathbf{s}, \mathbf{q}_Q) \end{pmatrix}^\top \quad \text{with} \quad v(\mathbf{s}, \mathbf{q}_i) = \frac{1}{N} \sum_{\forall \mathbf{p}_j \in \mathcal{P}} f(\mathbf{s}, \mathbf{p}_j, \mathbf{q}_i), \quad \mathbf{q}_i \in \mathcal{Q} \tag{12}$$

where N is the number of microphone pairs, Q is the size of $\mathcal{Q}$ and $f(\mathbf{s}, \mathbf{p}_j, \mathbf{q}_i) \in \mathcal{F}$ are the basis functions defined in Equation (10).

Vector $\mathbf{v}(\mathbf{s})$ can be intuitively seen as the ideal *SRP-PHAT* measurements that would be obtained for a single acoustic source located at position $\mathbf{s}$. If $\mathcal{Q}$ contains points with constant height, $\mathbf{v}(\mathbf{s})$ can be visualized as an image, composed as the sum of hyperbolic areas (one for each pair of microphones), intersecting at point $\mathbf{s}$ (see Figure 4). It must be remarked that $\mathbf{v}$ is normalized by definition, *i.e.*, $\max(\mathbf{v}(\mathbf{s})) = 1$.

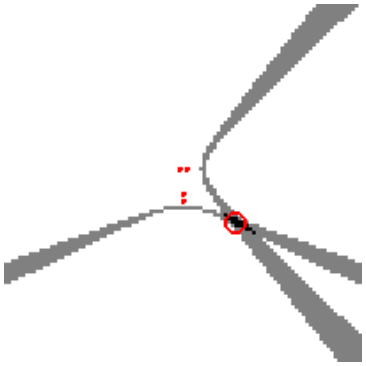
The proposed generative model consists of the following linear system:

$$\hat{\mathbf{y}} = \mathbf{M}\mathbf{x} \qquad \text{with} \quad \mathbf{M} = \begin{pmatrix} \mathbf{v}(\mathbf{s}_1) & \cdots & \mathbf{v}(\mathbf{s}_Q) \end{pmatrix} \qquad \mathbf{s}_i \in \mathcal{Q} \tag{13}$$

where $\mathbf{x} = (x_1, \cdots, x_Q)^\top$ is a vector of size Q, representing a numerical weight associated to each position considered in set $\mathcal{Q}$, where an acoustic source could be active. In fact, weight x_i corresponds

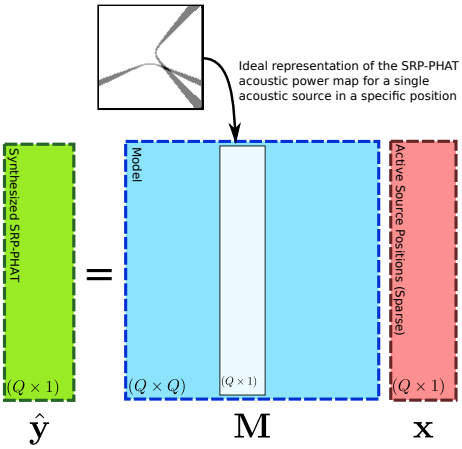
exactly to weight $\omega(\mathbf{s}_i)$ defined in Equation (9) up to a scale factor. In this case, $\mathbf{x}$ are the unknown parameters of the model.

Figure 4. Model content defined for a single active speaker located in the position of the red circle.



Matrix $\mathbf{M}$ is a $Q \times Q$ matrix whose columns are obtained using vector $\mathbf{v}$ defined at every $\mathbf{s} \in \mathcal{Q}$. Vector $\hat{\mathbf{y}}$ can be seen as the *SRP-PHAT* data synthesized by the proposed model as a function of weight vector $\mathbf{x}$. Figure 5 shows a graphical diagram of the proposed linear model.

Figure 5. Explicit matrix layout for the model proposal given by Equation (13).



Expanding the terms in Equation (13), vector $\hat{\mathbf{y}}$ is obtained as the following weighted sum of vectors:

$$\hat{\mathbf{y}} = x_1\mathbf{v}(\mathbf{s}_1) + \cdots + x_Q\mathbf{v}(\mathbf{s}_Q) \tag{14}$$

where it is explicitly seen that weight x_i directly affects the influence of vector $\mathbf{v}(\mathbf{s}_i)$ in the output vector $\hat{\mathbf{y}}$. Therefore, if vector $\mathbf{x}$ has high values around a single position $\mathbf{s}_i$, the resulting vector $\hat{\mathbf{y}}$ will have a maximum at $\mathbf{s}_i$, producing a *SRP-PHAT* image consistent with the model presented in the previous section. Nevertheless, as it was discussed in Section 3.3, it must be recalled that the hyperbolic model defined by Equation (10) is only a rough simplification of the real phenomenon, where noise,

reverberation and array directivity issues produce artifacts in the *SRP-PHAT* approximation that are not considered in the model. The consideration of these additional effects in the formulation of the basis functions can lead to improvements in the modeling ability of the proposed solution.

4. Model Fitting

This section explains how to use the linear model proposed in the previous section to fit real *SRP-PHAT* data. One of the main contribution of the paper is to show that as a result of model fitting, the performance of *SRP-PHAT* based localization techniques can be remarkably improved.

Suppose that vector $\mathbf{y}$ contains *SRP-PHAT* measurements (arranged in a column vector) obtained in a real scenario:

$$\mathbf{y} = \begin{pmatrix} srp(\mathbf{q}_1) & \cdots & srp(\mathbf{q}_Q) \end{pmatrix}^\top \qquad q_i \in \mathcal{Q} \tag{15}$$

with $srp(\mathbf{q}_i)$ defined in Equation (1).

Our aim is finding a vector $\mathbf{x}$ capable of explaining $\mathbf{y}$ using model $\mathbf{M}$. It is expected that $\mathbf{y}$ includes modeling errors, reverberation, array directivity effects, and noise, thus making the proposed model invalid for an exact representation of $\mathbf{y}$. Instead, the goal will be finding a vector $\mathbf{x}$ capable to *better* explain $\mathbf{y}$. The notion of which vector $\mathbf{x}$ is better at modeling $\mathbf{y}$ can be answered using optimization techniques.

The basic approach is then to solve the following optimization problem:

$$\min_{\mathbf{x}} \rho(\mathbf{y}, \hat{\mathbf{y}}) = \min_{\mathbf{x}} \rho(\mathbf{y}, \mathbf{M}\mathbf{x}) \tag{16}$$

where ρ is a metric measuring how different are the measurements $\mathbf{y}$ and the vector $\hat{\mathbf{y}}$ generated by the model (*i.e.*, $\mathbf{M}\mathbf{x}$ from Equation (13)). A straightforward and somehow natural choice for ρ is to use the Euclidean distance as a metric:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 \tag{17}$$

which yields to a linear least squares problem. If matrix $\mathbf{M}$ has full rank, the minimum of Equation (17) is unique and can be obtained in closed-form. Otherwise a regularized problem can be solved instead using Tikhonov regularization [45]. In either case, solving problem (17) represents a weak approach when the model $\mathbf{M}$ is not accurate enough to fit the data $\mathbf{y}$, which contains noise and effects that cannot be reproduced by the model.

The approach of this paper, and one of the basis of our contribution, is to include additional constraints into Equation (17) able to give meaningful answers for $\mathbf{x}$ with noisy measurements, and for relatively simple basis functions in the generative model. Two basic improvements of problem (17) are proposed and detailed next.

4.1. Adding Sparse Constraints

In this paper it is assumed that there is only a small number of simultaneous active acoustic sources inside the space defined by $\mathcal{Q}$, which is a reasonable assumption in the majority of scenarios considered. Given that values of $\mathbf{x}$ represent positions in which there is an active acoustic source, it is thus sensible to force $\mathbf{x}$ to have as many zeroes as possible. In the mathematical language that means to force the

vector $\mathbf{x}$ to be a *sparse* vector, in which the number of non-zero elements is limited. In the optimization scheme, making the $\mathbf{x}$ vector to be *as sparse as possible* is equivalent to forcing the l_0 norm of $\mathbf{x}$ to be minimum.

Finding the vector $\mathbf{x}$ that simultaneously reduces the error between the input data and the model and forces $\mathbf{x}$ to be as sparse as possible can be mathematically expressed as follows:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 < \eta \tag{18}$$

where η is a real value that bounds the amount of error and model mismatch that is admissible. Minimizing (18) is very difficult as the l_0 norm makes the problem highly non-linear, NP-Hard and non-convex. No practical method guarantees the global convergence in this case.

Sparse optimization methods have received remarkable attention from the scientific community. Despite its theoretical complexity, several methods and approximations have been proposed so far, and of special relevance are those methods based on using the l_1 norm as a convex relaxation of the l_0 norm [33,46]. This relaxation transforms (18) into the following:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{M}\mathbf{x}\|^2 < \gamma \tag{19}$$

where γ is an hyperparameter closely related to η in (18). Equivalently, problem (19) can be expressed in its Lagrangian form:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{20}$$

where λ is the Lagrange multiplier and has a direct relationship with γ.

Both Equations (19) and (20) are equivalent convex problems, in which convergence is guaranteed and can be solved in polynomial time.

The problem of finding a least squares estimation subject to a l_1 restriction has been independently presented and popularized under the names of *Least Absolute Shrinkage Selection Operator (LASSO)* [47] and *Basis Pursuit Denoising* [32], being object of intensive study. In the past few years numerous optimization methods have been proposed, some of them adapted to specific problems.

Additionally, several generic libraries and toolboxes implementing those methods have been developed and are being extensively used. The results shown in the paper have been generated using one of these libraries [48], using a truncated Newton interior-point method, described in [49].

Solving the relaxed problem (20) does not necessary imply finding the solution to the original l_0 problem. The closeness and validity of l_1 relaxations have been extensively studied [33]. In some problems, the structure of matrix $\mathbf{M}$ and the expected degree of sparsity in the solution can make l_1 relaxations to be exact. For general linear systems, as it is the case in this paper, where matrix $\mathbf{M}$ has no apparent structure, l_1 relaxation empirically tends to impose only approximate sparse solutions. This paper provides strong experimental evidence of the improvements obtained by imposing l_1 penalties, effectively making the solution $\mathbf{x}$ more sparse. Sparsity is a strong "prior" that helps to bias the solution $\mathbf{x}$ so that the effect of noise and model mismatches are properly attenuated.

4.2. Adding Subspace Filtering

Although sparsity is a well founded constraint and the l_1 relaxations are effective, the experimental results in Section 5 show that, given the current model, sparsity is not strong enough to cope with errors and model mismatches in real *SRP-PHAT* measurements so that additional strategies must be used to improve model fitting.

This section introduces a new constraint on the problem based on filtering out the part of the input signal $\mathbf{y}$ that is not reproducible using model $\mathbf{M}$.

First decompose $\mathbf{y}$ into two parts:

$$\mathbf{y} = \hat{\mathbf{y}} + \tilde{\mathbf{y}} = \mathbf{M}\mathbf{x} + \tilde{\mathbf{y}} \tag{21}$$

where $\hat{\mathbf{y}}$ is a term that can be explained exactly by the generative model (*i.e.*, there exists a vector $\mathbf{x}$ such that $\hat{\mathbf{y}} = \mathbf{M}\mathbf{x}$) and $\tilde{\mathbf{y}}$ represents the non-reproducible part of the signal (*i.e.*, $\tilde{\mathbf{y}} \neq \mathbf{M}\mathbf{x}$ for any vector $\mathbf{x}$). This section proposes to use subspace filtering to remove the non-reproducible part $\tilde{\mathbf{y}}$ from the input vector $\mathbf{y}$.

First, matrix $\mathbf{M}$ is expressed using *singular value decomposition* (SVD) as follows:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \tag{22}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of dimensions $Q \times Q$ and $\mathbf{\Sigma}$ is a semidefinite positive diagonal matrix of dimension $Q \times Q$. The diagonal elements of $\mathbf{\Sigma}$ are the singular values, sorted in descending order. Using singular values it is possible to know the amount of degrees of freedom available in the model by just looking how many non-zero singular values it has.

By identifying the number of zero singular values of $\mathbf{M}$, namely N_z, the SVD decomposition shown in Equation (22) can be expressed using the following sub-matrices:

$$\mathbf{M} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_0 \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^* \\ \mathbf{V}_0^* \end{pmatrix} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* \tag{23}$$

where $\mathbf{U}_0$ and $\mathbf{V}_0$ are $Q \times N_z$ matrices, $\mathbf{U}_1$ and $\mathbf{V}_1$ are of size $Q \times (Q - N_z)$ and $\mathbf{\Sigma}_1$ is a diagonal $(Q - N_z) \times (Q - N_z)$ matrix.

$\mathbf{U}_1$ and $\mathbf{U}_0$ are subspace projection matrices. Any nonzero vector $\mathbf{z}$ such that $\mathbf{U}_1^\top \mathbf{z} = 0$ is a vector that cannot be obtained using the model $\mathbf{M}$, *i.e.*, $\mathbf{z} \neq \mathbf{M}\mathbf{x}$ for any possible $\mathbf{x}$.

So, recalling that $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, both sides of the equality (21) can be multiplied by $\mathbf{U}^*$ with the following result:

$$\begin{pmatrix} \mathbf{U}_1^*\mathbf{y} \\ \mathbf{U}_0^*\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_1^*\mathbf{V}_1^* \\ \mathbf{0} \end{pmatrix} \mathbf{x} + \begin{pmatrix} \mathbf{U}_1^*\tilde{\mathbf{y}} = \mathbf{0} \\ \mathbf{U}_0^*\tilde{\mathbf{y}} \end{pmatrix} \tag{24}$$

By definition, if $\tilde{\mathbf{y}}$ cannot be expressed by the model, then its projection using matrix $\mathbf{U}_1^\top$ must be zero. Contrary, the projection into the kernel subspace represented by $\mathbf{U}_0$ is nonzero.

Therefore, in order to remove the dependence of $\tilde{\mathbf{y}}$, only the Mahalanobis distance of the upper part of system (24) is optimized, regularized with the l_1 term, and resulting into the problem (20) to become:

$$\min_{\mathbf{x}} \|\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^\top\mathbf{y} - V_1^*\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{25}$$

In practice, a small threshold ψ is used to decide if a singular value can be considered zero. Experiments are carried out in Section 5.5 to learn the value of parameter ψ from real *SRP-PHAT* data, which turns out to be an important parameter in practice. In order to give meaningful discrete values to ψ this paper uses the following ratio:

$$r(\psi) = \frac{\sum_{\lambda_j > \psi} \lambda_j}{\sum_{i=1}^{Q} \lambda_i} 100 \tag{26}$$

where $diag(\mathbf{\Sigma}) = (\lambda_1, \cdots, \lambda_Q)^\top$ are the singular values of $\mathbf{M}$. The meaning of Equation (26) is basically the percentage of Frobenius norm that $\mathbf{M}$ has lost after filtering out small singular values using ψ. By bounding the ratio with an *energy* threshold, namely $e_\psi \in [0\%, 100\%]$, which can be chosen easily with independence of scale factors (e.g., $e_\psi = 50\%$ means half of the energy in the model), the value of ψ can be chosen adequately as:

$$\min_{\psi} \qquad \text{s.t.} \qquad r(\psi) \leq e_\psi \tag{27}$$

In Section 5, the value of ψ is chosen by giving values to e_ψ using (27) afterwards.

After setting to zero all the N'_z singular values below threshold ψ, we can build new matrices $\mathbf{U'}_0$ and $\mathbf{V'}_0$ ($Q \times N'_z$), $\mathbf{U'}_1$ and $\mathbf{V'}_1$ ($Q \times (Q - N'_z)$) and $\mathbf{\Sigma'}_1$ ($(Q - N'_z) \times (Q - N'_z)$), for which the SVD decomposition (23) becomes:

$$\mathbf{M'} = \begin{pmatrix} \mathbf{U'}_1 & \mathbf{U'}_0 \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma'}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V'}_1^* \\ \mathbf{V'}_0^* \end{pmatrix} = \mathbf{U'}_1 \mathbf{\Sigma'}_1 \mathbf{V'}_1^* \tag{28}$$

and the optimization problem (25) becomes:

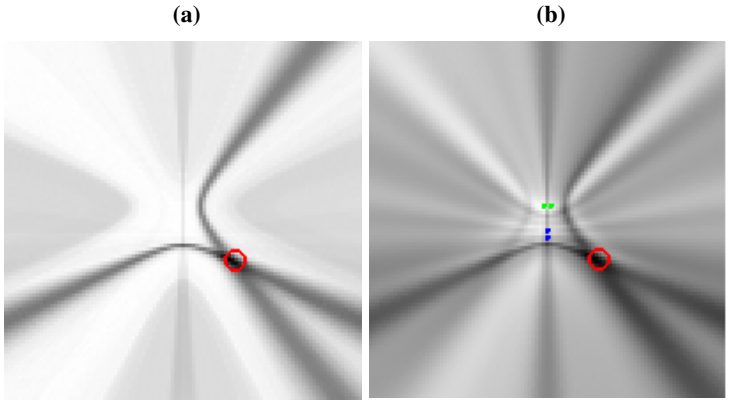
$$\min_{\mathbf{x}} \|\mathbf{\Sigma'}_1^{-1} \mathbf{U'}_1^\top \mathbf{y} - \mathbf{V'}_1^* \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{29}$$

4.3. Improving SRP-PHAT *with Model Fitting*

The main objective of the paper is to show that, as a result of the optimization methods proposed before, the solution $\mathbf{x}$ can be used to improve source localization, comparing with traditional approaches directly using *SRP-PHAT* measurements. The detection of local maxima in *SRP-PHAT* acoustic power maps is the standard way to retrieve the position of the acoustic source. This technique yields good results but is still prone to errors due to reverberation and noise and when the number of microphones is limited.

Our approach consists of replacing the original *SRP-PHAT* measurements $\mathbf{y}$ with those generated by the model solving the optimization (29), *i.e.*, $\hat{\mathbf{y}}' = \mathbf{M'}\mathbf{x'}$, where $\mathbf{M'}$ is obtained from Equation (28) and $\mathbf{x'}$ is the solution of Equation (29). Vector $\hat{\mathbf{y}}'$ can also be interpreted as a filtered/denoised version of $\mathbf{y}$ that is consistent with the proposed model. Figure 6 shows the acoustic power map described by the denoised vector $\hat{\mathbf{y}}'$ (Figure 6(a)) and the original *SRP-PHAT* acoustic power map $\mathbf{y}$ (Figure 6(b)). From the figure, it seems clear that the *denoising* effectively reduces the number of artifacts and unwanted effects exhibited by the original map, and the assumption is that this *denoised* version $\hat{\mathbf{y}}'$, if properly constrained during the optimization, is a better place to find local maxima truly representing active acoustic sources. In Section 5 the paper gives strong experimental indicators to support this idea.

Figure 6. Comparison between real *SRP-PHAT* power map and its denoised version. (a) Denoised acoustic power map described by $\hat{\mathbf{y}}'$; (b) Real *SRP-PHAT* acoustic power map described by $\mathbf{y}$.

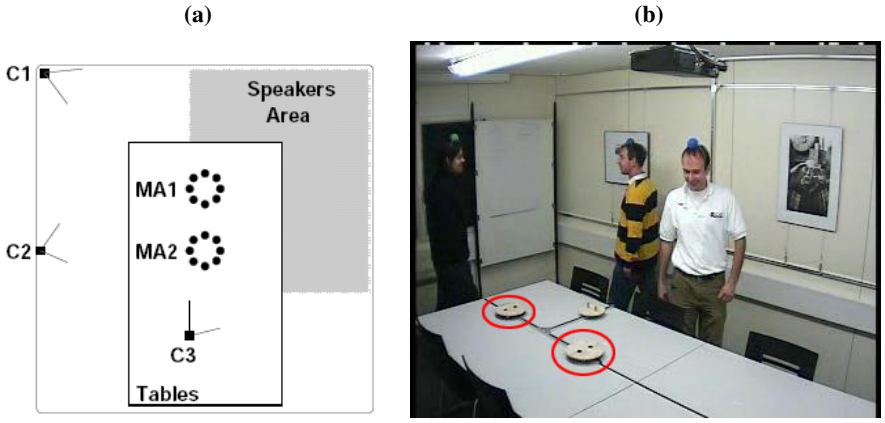


5. Experiments and Discussion

5.1. Experimental Setup

We have evaluated our proposal using the audio recordings of the AV16.3 database [50], an audio-visual corpus recorded in the *Smart Meeting Room* of the IDIAP research institute, in Switzerland.

Figure 7. Idiap Smart Meeting Room for AV16.3 recordings (a) Room layout showing the microphone positions in two circular arrays (MA1 and MA2), three cameras (C1, C2 and C3), and the L-shaped area for speaker locations in the recordings. (b) Sample of recorded video frame.



The IDIAP Meeting Room consists on a 8.2 m $\times$ 3.6 m $\times$ 2.4 m rectangular room containing a centrally located 4.8 m $\times$ 1.2 m rectangular table, on top of which two circular microphone arrays of 0.1 m radius are located, each composed by 8 microphones. The centers of the two arrays are separated by 0.8 m and the origin of coordinates is located in the middle point between the two arrays. Possible

speakers' locations are distributed along a L-shaped area around the table as seen in Figure 7(a). A detailed description of the meeting room can be found in [51].

The audio recordings are synchronously sampled at 16 KHz, and the complete database along with the corresponding annotation files containing the recordings ground truth is fully accessible on-line at [52]. It is composed by several sequences or recordings which range in the number of speakers involved and their activity. In this paper we will just focus on the single static speakers sequences, whose main characteristics are shown in Table 1. We will refer to the sequences as *seq01*, *seq02* and *seq03* for brevity.

Table 1. Characteristics of the audio sequences used in the experimental results.

Sequence name	speaker	Average speaker height* (m)	duration(s)	number of ground truth frames
seq01-1p-0000	male	54.3	208	2,248
seq02-1p-0000	female	62.5	171	2,411
seq03-1p-0000	male	70.3	220	2,636

* In the reference coordinate system.

Every audio sequence is assigned a corresponding annotation file containing the real ground truth positions (3D coordinates) of the speaker's mouth at every time frame in which that speaker was talking. The segmentation of acoustic frames with speech activity was first checked manually at certain time instances by a human operator in order to ensure its correctness, and later extended to cover the rest of recording time by means of interpolation techniques. The frame shift resolution was defined to be 40 ms.

5.2. Evaluation Metrics

Our localization algorithm yields a set of spatial coordinates $\mathbf{q}(t) = (x, y, z)^\top$ that are estimations of the actual speaker position, for every time frame t. These position estimates will be compared, by means of the Euclidean distance, to the ones labeled in a transcription file containing the real positions $\mathbf{s}(t)$ (*ground truth*), of the speaker.

We have decided to use the metrics developed under the CHIL project and described in their Evaluation Plan [53]. A complete description of the CHIL Evaluation strategies can be found at [53], but in this work we will only refer to the *Multiple Object Tracking Precision* ($MOTP$), calculated as the average localization error for all (N_T) acoustically active frames in the data set: $MOTP = \frac{\sum_{t=1}^{N_T} \|\mathbf{q}(t) - \mathbf{s}(t)\|}{N_T}$.

5.3. Evaluation Plan

We are evaluating our model in a 2D scenario, considering the acoustic power maps generated by *SRP-PHAT* at locations $\mathcal{Q}$ belonging to a plane 61 cm above the microphone array positions (this height roughly corresponds to the average height of the speaker positions in the AV16.3 sequences). Locations for *SRP-PHAT* data are calculated uniformly sampling $\mathcal{Q}$ in a 10 cm $\times$ 10 cm grid.

The procedure to generate the position estimations $\mathbf{q}(t)$ consists of searching for maximum values in vector $\hat{\mathbf{y}}'$ (calculated as described in Section 4.3) that could be seen as a *denoised* version of the original *SRP-PHAT* acoustic power map.

In the experimental results shown below, we are assessing the performance of our proposal in terms of:

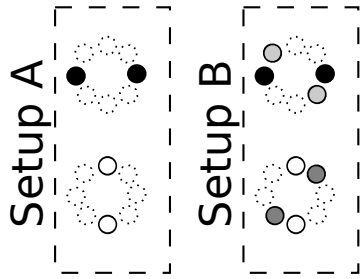
- Optimization parameters: We will provide results depending on the two main tunable parameters of the optimization algorithms used, namely λ and e_ψ.

  The estimation of the optimal values for this parameters will be done on an independent data set (training set) and applied to unseen data in the evaluation stage (test set).

- Sensor array configuration: In this work, we are using a simple microphone array configuration, aimed at evaluating our proposal in a resource-restricted environment. In order to do so, we are using 4 or 8 microphones (out of the 16 available in the AV16.3 data set), grouped in two or four microphone pairs to generate the baseline *SRP-PHAT* acoustic maps. The selected microphone pairs configurations are shown in Figure 8, in which microphones with the same color are considered as belonging to the same microphone pair. Given that the microphone separation for each microphone pair is 20 cm, we will violate spatial aliasing requirements, considering the signal bandwidth. Fortunately, when using *SRP-PHAT*, the use of more than one microphone pair alleviates this problem, as side lobes are different for each pair, and thus their effects are partially compensated.

- Acoustic frame size: We will provide results depending on the length of the acoustic frame, for 80, 160 and 320 ms, to precisely assess to what extent the improvements are consistent with varying acoustic time resolutions.

Figure 8. Microphone pairs setups used in the experiments (microphones with the same color belong to the same pair).



The baseline we are comparing with will be the results of directly searching the maximum of the *SRP-PHAT* acoustic power map. The position of this maximum will correspond to the most probable source location.

Comparisons will specifically consider the relative improvement in $MOTP$, defined as $\Delta_r^{MOTP} = \frac{MOTP_{baseline} - MOTP_{proposal}}{MOTP_{baseline}}$.

Our main interest is assessing whether the results and improvements are consistent across different conditions. After describing the baseline results (in Section 5.4) and in order to evaluate the generalization capability of the proposed methods, we will address an initial study using sequence *seq01* as the *training set* (in Section 5.5). From this study, we will decide on the optimal values of the tunable parameters used in the optimization process (those leading to the best results), and then use them to provide final performance and improvement results on the *test sets*, namely *seq02* and *seq03* (in Section 5.6). This evaluation plan ensures adequate independence and variability between train and test sets, with different speakers in all sequences (also differing in gender and height).

In all cases were appropriate, we will include references to statistical confidence values for a 95% confidence level, to adequately assess whether the improvements are statistically significant.

5.4. Baseline Results

Tables 2 and 3 show the baseline results using the standard *SRP-PHAT* algorithm for all sequences and different frame sizes, and the two microphone setups of Figure 8.

Table 2. Baseline $MOTP(m)$ results for all sequences, different frame sizes and microphone setup A.

		80 ms	160 ms	320 ms
seq01	$MOTP$	1.02 ± 0.03	0.91 ± 0.03	0.83 ± 0.03
seq02	$MOTP$	0.96 ± 0.03	0.84 ± 0.03	0.77 ± 0.02
seq03	$MOTP$	0.90 ± 0.03	0.77 ± 0.03	0.69 ± 0.03

Table 3. Baseline $MOTP(m)$ results for all sequences, different frame sizes and microphone setup B.

		80 ms	160 ms	320 ms
seq01	$MOTP$	0.87 ± 0.03	0.74 ± 0.03	0.62 ± 0.02
seq02	$MOTP$	0.73 ± 0.02	0.62 ± 0.02	0.56 ± 0.02
seq03	$MOTP$	0.71 ± 0.02	0.59 ± 0.02	0.50 ± 0.01

The main conclusions for the baseline results are:

- The performance obtained is reasonable if we take into account that only two or four microphone pairs are used. Best $MOTP$ values are around 50 cm.
- Performance improves as the frame size increases, as expected, given that longer frames lead to better estimations of the correlation functions.
- Adding an additional microphone pair in setup B as compared with setup A also leads to performance improvements as expected.

5.5. Evaluation of the Sensitivity to λ and e_ψ Values

The proposed model fitting strategies heavily depend on the estimation of adequate values for both λ and e_ψ (as they are the parameters controlling the optimization process), so that a detailed study on the sensitivity of the performance with variations in these parameter values is mandatory.

λ expresses the relative importance of the sparse constraints applied in the optimization problems (20), (25) and (29), so that the higher its value becomes, the sparser the solution will be. In the l_1 optimization software used [48], it is required that $\lambda < \lambda_{max}$ being λ_{max} dependent on both the model and the input data [49]. In the results shown, the hyperparameter is represented normalized with respect to the calculated λ_{max}: $\lambda_{norm} = \lambda/\lambda_{max}$, as described in [49].

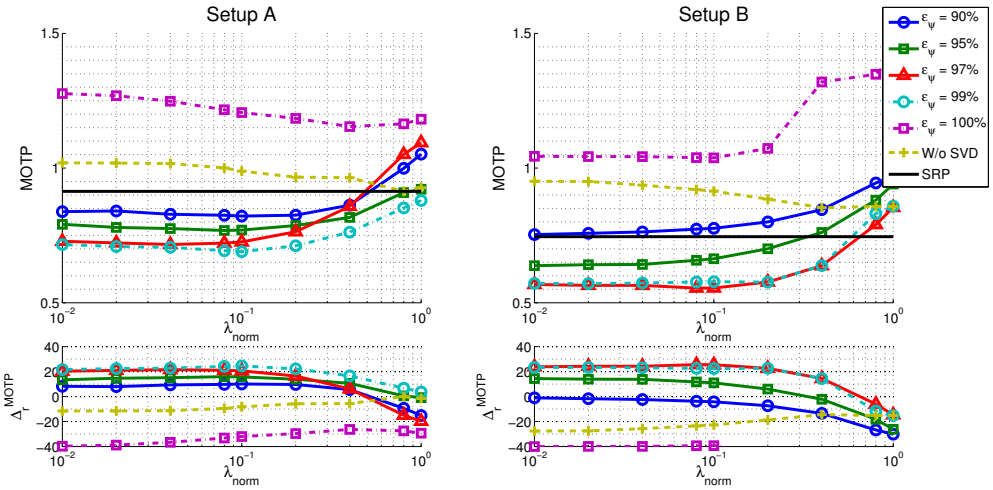
The energy threshold e_ψ used in the subspace filtering strategy described by Equation (27) decides the size of the model that is not able to adequately *explain* the input signal.

To decide on the optimal λ_{norm} and e_ψ to be used, we will select the values that achieve the best result in terms of $MOTP$, for every microphone setup and frame size.

In the upper part of Figure 9, we show the evolution of the $MOTP$ quality metric as a function of λ_{norm} and the energy value e_ψ, for both microphone setups, evaluating the training sequence *seq01*, with a frame size of 160 ms, as an example. The horizontal black trace show the baseline results for the *SRP-PHAT* algorithm (obviously independent of λ_{norm} and e_ψ). In the lower part of Figure 9 the evolution of the relative improvements in $MOTP$ are shown.

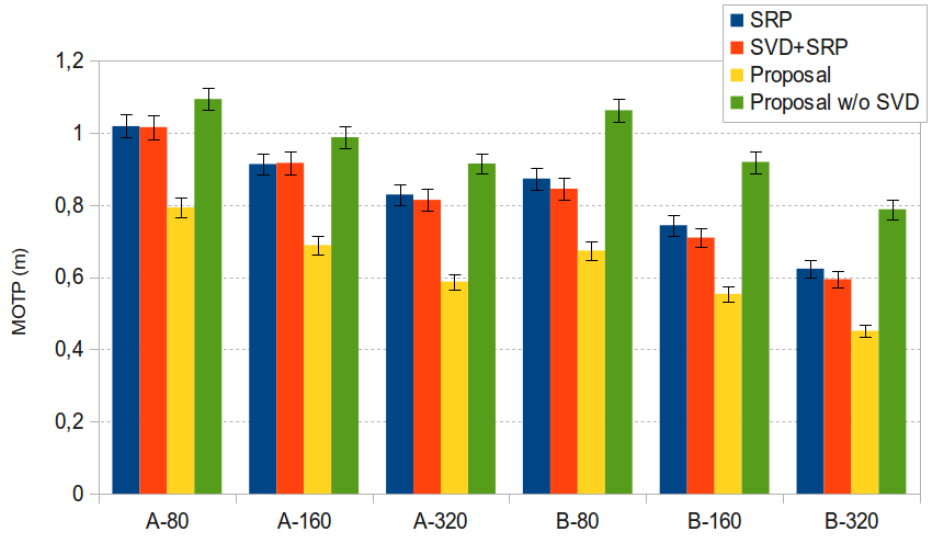
Additionally, and in order to evaluate the effectiveness of the subspace filtering step, we ran an experiment in which only the optimization with sparse constraints described in Equation (20) is applied (*i.e.*, our proposal without using subspace filtering). The results are shown in the "W/o SVD" trace of Figure 9.

Figure 9. Optimization results for $MOTP$ and relative improvements as a function of λ_{norm} and e_ψ, for microphone setups A and B on sequence *seq01*. The black trace is the baseline *SRP-PHAT* system.



In Figure 10 we show the best $MOTP$ results for sequence *seq01* for both microphone setups and all frame sizes, with 95% confidence intervals. Data includes results for the baseline *SRP-PHAT* results ("SRP" in the legend, blue bars), for our proposal ("Proposal" in the legend, yellow bars), and for our proposal without applying the SVD step ("Proposal w/o SVD" in the legend, green bars) (the orange bar ("SVD+SRP" in the legend) refers to results that will be discussed later).

Figure 10. Best $MOTP$ results for sequence *seq01* for both microphone setups (A, B) and all frame sizes (80, 160 and 320 ms.), with 95% confidence intervals.



From this, we can conclude that, for adequate values of the optimization tuning parameters:

- Our proposal is able to improve the *SRP-PHAT* results with statistically significant relative improvements of up to almost 25%, with consistent improvements for a wide range of λ_{norm} values.
- Microphone setups have a similar impact in the relative performance improvements. The improvements for setups A and B are 24.6% and 25.6%, respectively.
- In what respect to the dependency of the best results with λ_{norm} (once selected the optimal e_ψ), both microphone setups show a desirable behavior, achieving a reasonably clear optimal area for a wide range of parameter values.
- Using either the model with sparse constraints (*i.e.*, "Proposal w/o SVD") or SVD without actually filtering (*i.e.*, $e_\psi = 100\%$) is giving worse localization results than the *SRP-PHAT* baseline algorithm. It thus seems that fitting the complete model to data is not making any progress even if sparse constraints are included. The explanation of this phenomenon was partially advanced in Section 4.2 but it needs some additional justification. The model that is proposed in this paper is not able to explain every *SRP-PHAT* map (*i.e.*, matrix $\mathbf{M}$ is rank-deficient). When using any of the optimization strategies proposed in the paper, the position of speakers is the result of looking

at local maxima in the *SRP-PHAT* map reproduced through the model. Therefore, in theory, the results must not be necessarily equal to the baseline algorithm, even if subspace filtering is removed, or the l_1 term is not having strong influence. Empirical data tell us that in these cases, localization results can be in fact worse than the baseline. The main result of the paper is to show through experiments that statistically significant improvements can be reached using a specific combination of subspace filtering and sparse constraints. In these cases the model is able to adequately filter the effects of noise and reverberation in the *SRP-PHAT* map, giving a cleaner image about the real position of the speaker.

Table 4. Relative improvements of $MOTP(m)$ for sequence *seq01*, including the values of the optimal parameters, estimated per microphone setup and per frame size.

		80 ms	160 ms	320 ms
setup A	Δ_r^{MOTP}	22.1%	24.6%	29.1%
	$\lambda_{norm}^{optimal}$	0.1	0.1	0.1
	$e_{\psi}^{optimal}$	99%	99%	99%
setup B	Δ_r^{MOTP}	22.9%	25.6%	27.6%
	$\lambda_{norm}^{optimal}$	0.04	0.08	0.1
	$e_{\psi}^{optimal}$	97%	97%	97%

Table 4 shows the highest relative improvements obtained for sequence *seq01* and the optimal values of the parameters found to achieve these best results (namely $\lambda_{norm}^{optimal}$ and $e_{\psi}^{optimal}$). The table shows how the maximum improvement is high and consistent along different frame sizes and microphone setups. Improvements in $MOTP$ clearly increase as the frame size increases.

Table 5. Relative improvements of $MOTP(m)$ for sequence *seq01* and microphone setup B, using different values for the optimization parameters.

		80ms	160ms	320ms
setup B $e_{\psi}^{optimal-B} = 97\%$	Δ_r^{MOTP}	22.9%	24.2%	26.7%
	λ_{norm}	0.04	0.04	0.04
setup B $e_{\psi}^{optimal-B} = 97\%$	Δ_r^{MOTP}	22.1%	25.6%	27.2%
	λ_{norm}	0.08	0.08	0.08
setup B $e_{\psi}^{optimal-B} = 97\%$	Δ_r^{MOTP}	22.2%	25.3%	27.6%
	λ_{norm}	0.1	0.1	0.1
setup B $\lambda_{norm}^{optimal-A} = 0.1$ $e_{\psi}^{optimal-A} = 99\%$	Δ_r^{MOTP}	21.4%	22.6%	24.3%

Interestingly, the optimal values for the parameters controlling the optimization process are identical for all frame sizes in the setup A ($\lambda_{norm}^{optimal-A} = 0.1$ and $e_{\psi}^{optimal-A} = 99\%$). This seems not to be the case for setup B, in which $e_{\psi}^{optimal-B} = 97\%$ in all cases, but $\lambda_{norm}^{optimal-B}$ values varies for different frame sizes. However, even in this case, the improvements are stable for a wide range of parameter values as

can be seen in the first three rows of Table 5, where the relative improvements have been calculated for different values of λ_{norm} (0.04, 0.08 and 0.1), setting $e_\psi = e_\psi^{optimal-B} = 97\%$.

From Table 4 it also seems that the optimal values of the parameters are dependent on the microphone setup used, as both $\lambda_{norm}^{optimal}$ and $e_\psi^{optimal}$ are different for setups A and B. A more detailed evaluation shows that, again, the improvements are stable even when we use the optimal values estimated for setup A ($\lambda_{norm}^{optimal-A} = 0.1$ and $e_\psi^{optimal-A} = 99\%$), in the optimization process for setup B data, as it can be seen in the last row of Table 5.

An additional way of visually assessing to what extent the results of the optimal values for the optimization parameters are consistent for different situations is plotting a surface map of $MOTP$ versus variations on λ_{norm} and e_ψ and making a comparison. For example, Figures 11 and 12 show this *optimization map* for microphone setups A and B respectively, using sequence *seq01*. In these maps, the optimal points for each evaluation are represented with a *circle* for *seq01* and setup A, and with a *triangle* for *seq01* and setup A. The maps show a *similar* structure for the optimal region in both microphone setups, supporting the idea that the optimal optimization parameters do not heavily depend on changes of the experimental conditions. Moreover, in the cases for microphone setup B, where the optimal points (triangles) seem not to be close to the optimal points of setup A (circles), it can be seen that these positions *belong* to an area with roughly the same $MOTP$ level (the area can be recognized as a *flat* optimal region).

The main conclusion of these experiments is that, for the given experimental setup, our proposal is able to clearly outperform the standard *SRP-PHAT* results. The statistically significant relative improvements roughly vary between 22% and 30%, and, what is more important, with little sensitivity to the optimization parameters selected when changing the microphone setup and the frame size used (once the optimal parameters have been estimated for the training data).

Figure 11. Optimization map for microphone setup A on sequence *seq01*. The circle is the position of the best parameter combination.

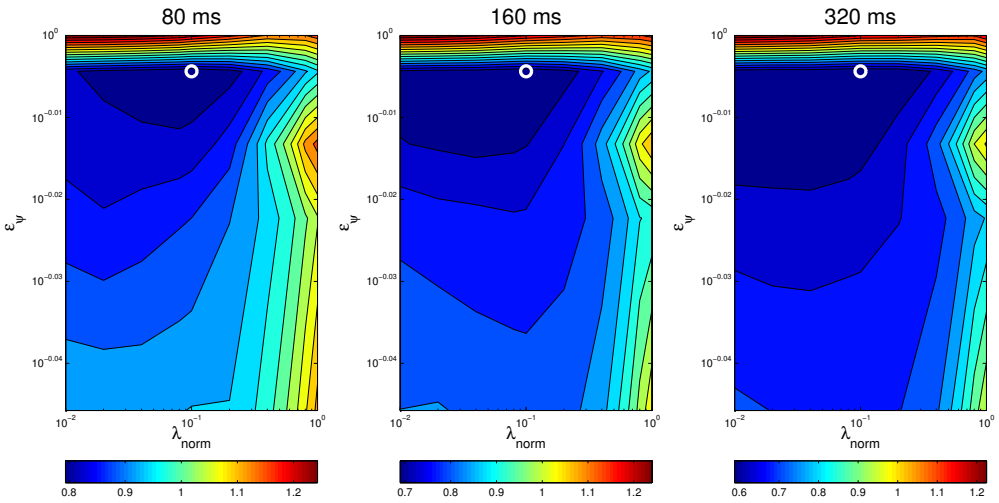
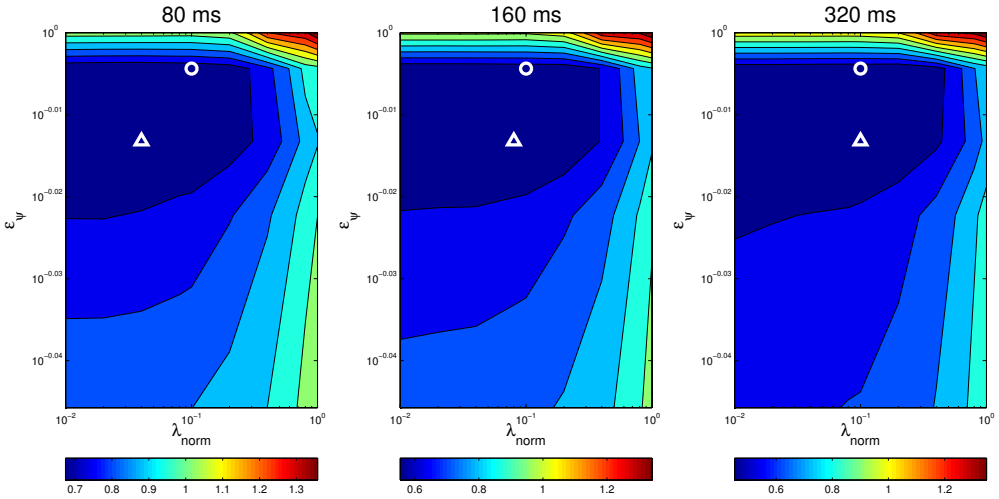


Figure 12. Optimization map for microphone setup B on sequence *seq01*. The circle is the position of the best parameter combination in *seq01* calculated for setup A and the triangle is the position of the best parameter combination in *seq01* calculated for setup B.



To further evaluate the contribution of the subspace filtering strategy, we ran an experiment in which we applied the subspace filtering to the original *SRP-PHAT* data, that is, projecting the *SRP-PHAT* acoustic power map on the span of model $\mathbf{M}'$ obtained from (28). This projection generates a new filtered power map, calculated as $\mathbf{y}^\star = \mathbf{U}'\mathbf{U}'^{\top}_{1}\mathbf{y}$. The results applying this transformation are given in the orange bars of Figures 10 and 15, referred to as "SVD+SRP". In these figures, we can see that SRP+SVD also outperforms SRP, although the differences are not statistically significant.

5.6. Evaluation on the Test Set

The evaluation carried out in the previous section only addresses the estimation of the optimal parameters for a single *training* sequence and the proposal evaluation on this same data set (*seq01*). We still need to assess whether the optimal values estimated for the *training* data set are able to achieve good results when using different sequences. As stated above, we are using *seq02* and *seq03* as independent *test sets*.

Figures 13 and 14 show the *optimization maps* for all sequences, frame size 160 ms, and microphone setups A and B, respectively. The *cross* is located in the optimal point for each sequence and setup A, and the *diamond* is located in the optimal point for each sequence and setup B. It can be seen that, again, the *structure* of the optimal regions are reasonably similar, thus suggesting that the optimal values for the optimization parameters estimated in the training set will also achieve good results in the test sets. The position of the optimal points in each map also belong to the same *flat* optimal region.

Figure 15 shows the best $MOTP$ results for sequences *seq02* and *seq03* for both microphone setups and all frame sizes, with 95% confidence intervals (using the optimal parameter values estimated for the training sequence *seq01*).

Figure 13. Optimization map for microphone setup A on all sequences, evaluating $MOTP$ and frame size 160 ms. The circle is the position of the best parameter combination calculated for sequence *seq01* and the cross is the best position calculated for each sequence.

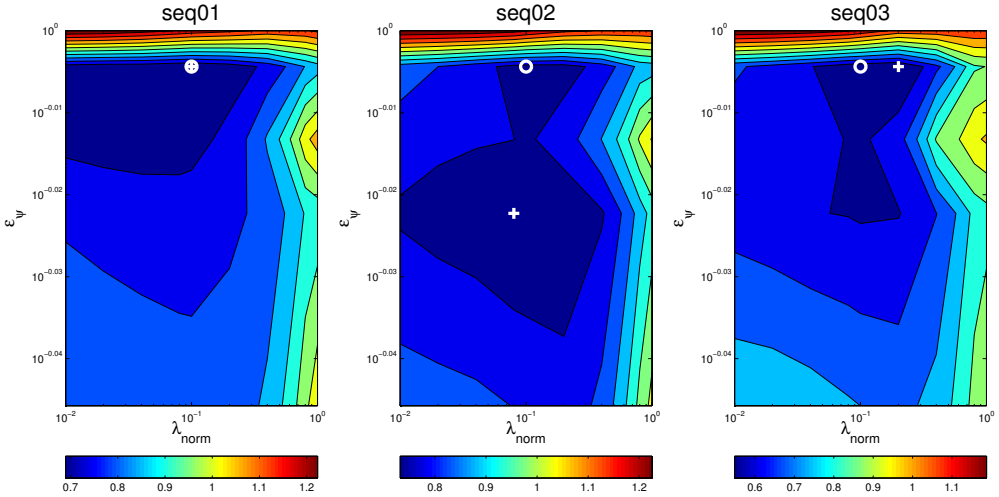


Figure 14. Optimization map for microphone setup B on all sequences, evaluating $MOTP$ and frame size 160 ms. The circle is the position of the best parameter combination calculated for sequence *seq01* with setup A,the triangle is the position of the best parameter combination in *seq01* with setup B and the diamond is the best position calculated for each sequence.

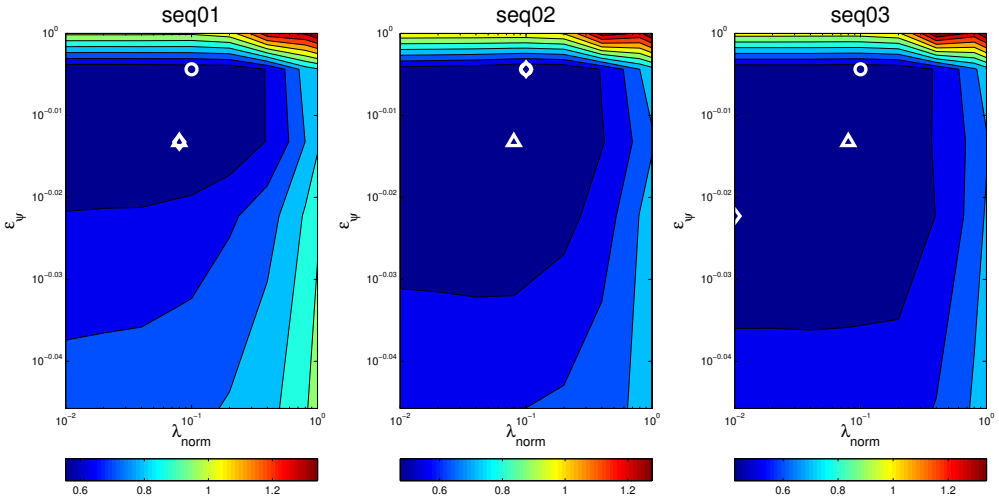


Figure 15. Best $MOTP$ results for both microphone setups (A, B) and all frame sizes (80, 160 and 320 ms), with 95% confidence intervals, (a) for sequence *seq02* and (b) for sequence *seq03*.

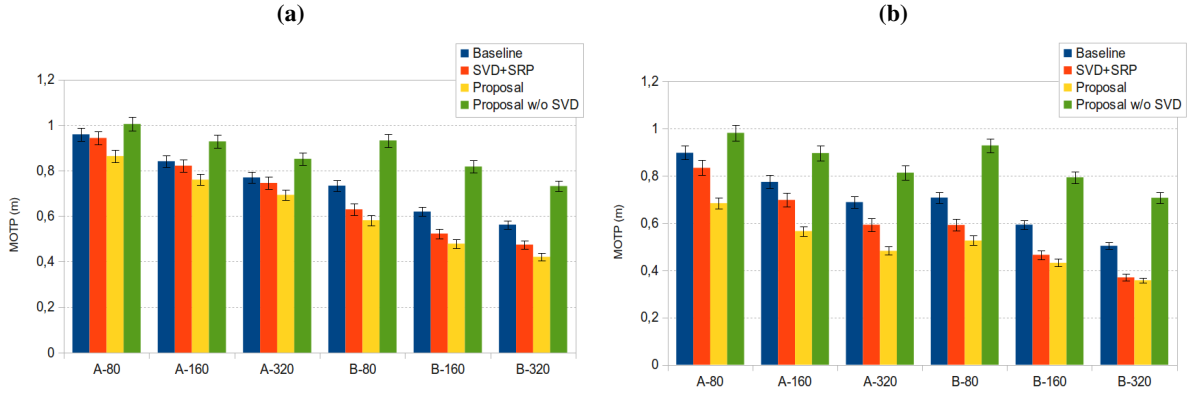


Table 6. Relative improvements of $MOTP(m)$ for sequence *seq02*, using the optimal parameter values estimated for sequence *seq01*.

		80 ms	160 ms	320 ms
setup A	Δ_r^{MOTP}	10.0%	9.6%	9.9%
	λ_{norm}	0.1	0.1	0.1
	e_ψ	99%	99%	99%
setup B	Δ_r^{MOTP}	20.7%	22.9%	25.1%
	λ_{norm}	0.04	0.08	0.1
	e_ψ	97%	97%	97%

Table 7. Relative improvements of $MOTP(m)$ for sequence *seq03*, using the optimal parameter values estimated for sequence *seq01*.

		80 ms	160 ms	320 ms
setup A	Δ_r^{MOTP}	23.8%	26.9%	29.9%
	λ_{norm}	0.1	0.1	0.1
	e_ψ	99%	99%	99%
setup B	Δ_r^{MOTP}	25.7%	27.3%	29.0%
	λ_{norm}	0.04	0.08	0.1
	e_ψ	97%	97%	97%

Tables 6 and 7 show the relative improvements achieved when evaluating sequences *seq02* and *seq03* for both microphone setups, also using the optimal parameter values for sequence *seq01*. As expected, the relative improvements are in the range of those obtained for sequence *seq01*, except for sequence *seq02* and microphone setup A (with lower improvements of around 10%). Our hypothesis is that the fact that this is a female speaker imposes significant differences in the speech signals, thus modifying the

correlation functions used in the input data, and posing additional difficulties to the optimization process when only two microphone pairs are used. Nevertheless, this will have to be evaluated in future work.

Apart from the case of *seq02* with setup A, the improvements are relevant and statistically significant, roughly varying between 20% and 30%. These achievements also show little sensitivity to the optimization parameters selected, in spite of the fact that we are additionally dealing with different speakers.

6. Conclusions and Future Work

This paper has proposed a novel method to localize active acoustic sources in a room equipped with sensor arrays. Two main contributions can be highlighted: First, a simple but very promising generative linear model is proposed to explain *SRP-PHAT* data taken from any geometrical combination of microphone arrays. The model simply reflects the geometry of three-dimensional points sharing common difference of time-of-arrival between each microphone pair. This model is independent of the spectrum properties of the signals emitted by the source and can be easily computed in practice. Second, this paper shows, using convincing experiments based on publicly available data, that such a simple model can be used to fit real *SRP-PHAT* data that is usually very noisy and has many unmodeled effects (such as reverberation in the scene). Fitting the model is done by imposing two constraints. The first one is forcing the model parameters to be sparse, as acoustic sources cannot be densely distributed in a typical environment. The second constraint simply removes the part of the measurements that is not exactly reproducible by the model. In the light of the experimental results, these two constraints in combination are the real key of the paper, notably improving the performance of state-of-the-art localization methods based on *SRP-PHAT*. It is also worth mentioning that all algorithms and experiments proposed in the paper are very easy to reproduce.

In future works the performance of this approach must be thoroughly validated in rooms with multiple speakers and using the whole three-dimensional set of spatial positions. Immediate improvements should cover all issues commented in Section 3.3. That means to propose basis functions in the model that take into account additional factors, such as the spectral content of the acoustic sources, directivity pattern effects in the microphone arrays, and also adding geometric information that would help to predict reverberation effects. The authors believe that improvements in the model may yield remarkable improvements in the localization accuracy in real world scenarios.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Innovation under projects VISNU (ref. TIN2009-08984) and SDTEAM-UAH (ref. TIN2008-06856-C05-05). We also thank the anonymous reviewers for their constructive feedback and helpful suggestions.

References

1. Weiser, M. Some computer science issues in ubiquitous computing. *Commun. ACM* **1993**, *36*, 75–84.

2. Pentland, A. Smart rooms. *Sci. Am.* **1996**, *274*, 54–62.
3. Lee, J.; Hashimoto, H. Intelligent space concept and contents. *Adv. Robot.* **2002**, *16*, 265–280.
4. Fleuret, F.; Berclaz, J.; Lengagne, R.; Fua, P. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 267–282.
5. Marron-Romera, M.; Garcia, J.; Sotelo, M.; Pizarro, D.; Mazo, M.; Cañas, J.; Losada, C.; Marcos, A. Stereo vision tracking of multiple objects in complex indoor environments. *Sensors* **2010**, *10*, 8865–8887.
6. Pizarro, D.; Mazo, M.; Santiso, E.; Marron, M.; Jimenez, D.; Cobreces, S.; Losada, C. Localization of mobile robots using odometry and an external vision sensor. *Sensors* **2010**, *10*, 3655–3680.
7. Lowe, D. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
8. Brandstein, M.S.; Silverman, H.F. A practical methodology for speech source localization with microphone arrays. *Comput. Speech Lang.* **1997**, *11*, 91–126.
9. DiBiase, J.; Silverman, H.; Brandstein, M. Robust localization in reverberant rooms. In *Microphone Arrays: Signal Processing Techniques and Applications*,1st ed.; Brandstein, M.S., Ward, D.B., Eds.; Springer-Verlag: New York, NY, USA, 2001; pp. 157–180.
10. Waibel, A.; Stiefelhagen, R. *Computers in the Human Interaction Loop*, 2nd ed.; Springer: London, UK, 2009.
11. DiBiase, J. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. Ph.D. Thesis, Brown University, Providence, RI, USA, 2000.
12. Gillette, M.; Silverman, H. A Linear Closed-Form Algorithm for Source Localization from Time-Differences of Arrival. *IEEE Signal Process. Lett.* **2008**, *15*, 1–4.
13. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327.
14. Zhang, C.; Florencio, D.; Zhang, Z. Why does PHAT work well in low noise, reverberative environments? In *Proceedings of ICASSP 2008 on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 2565–2568.
15. Dmochowski, J.P.; Benesty, J. Steered Beamforming Approaches for Acoustic Source Localization. In *Speech Processing in Modern Communication*, 1st ed.; Cohen, I., Benesty, J., Gannot, S., Eds.; Springer-Verlag: Berlin/Heidelberg, Germany, 2010; Volume 3, pp. 307–337.
16. Omologo, M.; Svaizer, P. Use of The Cross-Power-Spectrum Phase in Acoustic Event Location. *IEEE Trans. Speech Audio Process.* **1993**, *5*, 288–292.
17. Dmochowski, J.; Benesty, J.; Affes, S. A Generalized Steered Response Power Method for Computationally Viable Source Localization. *IEEE Trans. Speech Audio Process.* **2007**, *15*, 2510–2526.
18. Badali, A.; Valin, J.M.; Michaud, F.; Aarabi, P. Evaluating real-time audio localization algorithms for artificial audition in robotics. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, USA, 10–15 October 2009; pp. 2033–2038.

19. Do, H.; Silverman, H. SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data. In *Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, TX, USA, 14–19 March 2010; pp. 125–128.
20. Cobos, M.; Marti, A.; Lopez, J. A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. *IEEE Signal Proc. Lett.* **2011**, *18*, 71–74.
21. Butko, T.; Gonzalez Pla, F.; Segura Perales, C.; Nadeu Camprubí, C.; Hernando Pericás, F.J. Two-source acoustic event detection and localization: Online implementation in a smart-room. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO'11)*, Barcelona, Spain, 29 August–2 September 2011; pp. 1317–1321.
22. Habets, E.A.P.; Benesty, J.; Gannot, S.; Cohen, I. The MVDR Beamformer for Speech Enhancement. In *Speech Processing in Modern Communication*; Cohen, I., Benesty, J., Gannot, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 3, pp. 225–254.
23. Zhang, C.; Zhang, Z.; Florencio, D. Maximum Likelihood Sound Source Localization for Multiple Directional Microphones. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, 15–20 April 2007; Volume 1, pp. I-125–I-128.
24. Zhang, C.; Florencio, D.; Ba, D.; Zhang, Z. Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings. *IEEE Trans. Multimed.* **2008**, *10*, 538–548.
25. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antenn. Propag.* **1986**, *34*, 276–280.
26. Baraniuk, R. Compressive sensing [lecture notes]. *IEEE Signal Process. Mag.* **2007**, *24*, 118–121.
27. Candes, E. The restricted isometry property and its implications for compressed sensing. *C. R. Math.* **2008**, *346*, 589–592.
28. Rao, B.; Kreutz-Delgado, K. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.* **1999**, *47*, 187–200.
29. Davis, G.; Mallat, S.; Avellaneda, M. Adaptive greedy approximations. *Constr. Approx.* **1997**, *13*, 57–98.
30. Temlyakov, V. Nonlinear methods of approximation. *Found. Comput. Math.* **2003**, *3*, 33–107.
31. Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theor.* **2004**, *50*, 2231–2242.
32. Chen, S.; Donoho, D.; Saunders, M. Atomic decomposition by basis pursuit. *SIAM Rev.* **2001**, *43*, 129–159.
33. Tropp, J. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theor.* **2006**, *52*, 1030–1051.
34. Claerbout, J.; Muir, F. Robust modeling with erratic data. *Geophysics* **1973**, *38*, 826.
35. Malioutov, D.; Cetin, M.; Willsky, A. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Process.* **2005**, *53*, 3010–3022 .
36. Sun, K.; Liu, Y.; Meng, H.; Wang, X. Adaptive Sparse Representation for Source Localization with Gain/Phase Errors. *Sensors* **2011**, *11*, 4780–4793.

37. Ba, D.; Ribeiro, F.; Zhang, C.; Florêncio, D. L1 regularized room modeling with compact microphone arrays. In *Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, 14–19 March 2010; pp. 157–160.
38. Ribeiro, F.; Ba, D.; Zhang, C.; Floêncio, D. Turning enemies into friends: Using reflections to improve sound source localization. In *Proceedings of 2010 IEEE International Conference on Multimedia and Expo (ICME)*, Singapore, 19–23 July 2010; pp. 731–736.
39. Chardon, G.; Daudet, L. Narrowband source localization in an unknown reverberant environment using wavefield sparse decomposition. In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 25–30 March 2012; pp. 9–12.
40. Meuse, P.; Silverman, H. Characterization of talker radiation pattern using a microphone array. In *Proceedings of 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, 19–22 April 1994; pp. 257–260.
41. Chu, W.; Warnock, A. *Detailed Directivity of Sound Fields Around Human Talkers*; Research Report; Institute for Research in Construction: Ottawa, ON, Canada, 2002.
42. Wabnitz, A.; Epain, N.; Jin, C.T.; van Schaik, A. Room acoustics simulation for multichannel microphone arrays. In *Proceedings of the International Symposium on Room Acoustics*, Melbourne, Australia, 29–31 August 2010.
43. Kowalczyk, K.; van Walstijn, M. Room acoustics simulation using 3-D compact explicit FDTD schemes. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 34–46.
44. Ziomek, L.J. *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*; CRC Press: Boca Raton, FL, USA, 1995.
45. Tikhonov, A.; Arsenin, V.; John, F. *Solutions of Ill-Posed Problems*; Vh Winston: Washington, DC, USA, 1977.
46. Tropp, J.A. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Process.* **2006**, *86*, 589–602.
47. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **1996**, *58*, 267–288.
48. Koh, K.; Kim, S.; Boyd, S. l1_ls: A Matlab Solver for Large-Scale l1-Regularized Least Squares Problems. Available online: http://www.stanford.edu/ boyd/l1_ls/ (accessed on 11 October 2012).
49. Kim, S.; Koh, K.; Lustig, M.; Boyd, S.; Gorinevsky, D. An interior-point method for large-scale l1-regularized least squares. *IEEE J. Sel. Top. Signal Process.* **2007**, *1*, 606–617.
50. Lathoud, G.; Odobez, J.M.; Gatica-Perez, D. AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking. In *Proceedings of First International Workshop on Machine Learning for Multimodal Interaction*, Martigny, Switzerland, 21–23 June 2004.
51. Moore, D.C. *The IDIAP Smart Meeting Room*; Technical Report; IDIAP Research Institute: Martigny, Switzerland, 2004.
52. Lathoud, G. AV16.3 Dataset. Available online: http://www.idiap.ch/dataset/av16-3/ (accessed on 11 October 2012).

53. Mostefa, D.; Garcia, M.; Bernardin, K.; Stiefelhagen, R.; McDonough, J.; Voit, M.; Omologo, M.; Marques, F.; Ekenel, H.; Pnevmatikakis, A. *Clear Evaluation Plan*; Technical Report; Document CHIL-CLEAR-V1.1 Available online: http://www.clear-evaluation.org/clear06/downloads/chil-clear-v1.1-2006-02-21.pdf (accessed on 11 October 2012).

Chapter 4

Journal Paper on Modeling of the PHAT Filtering Effects: *Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios*

Publication reference:

- J. Velasco, C. J. Martín-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Processing*, vol. 119, pp. 209 – 228, 2016

Signal Processing 119 (2016) 209–228

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios

Jose Velasco [a], Carlos J. Martín-Arguedas [a], Javier Macias-Guarasa [a,*], Daniel Pizarro [b], Manuel Mazo [a]

[a] Department of Electronics, University of Alcalá, Alcalá de Henares, Spain
[b] ALCoV-ISIT, Université d' Auvergne, 63001 Clermont-Ferrand, France

ARTICLE INFO

Article history:
Received 12 October 2014
Received in revised form
4 April 2015
Accepted 2 August 2015
Available online 20 August 2015

Keywords:
Microphone arrays
Steered Response Power (SRP)
Generalized Cross-Correlation (GCC)
Phase Transform (PHAT)
Acoustic Source Location (ASL)

ABSTRACT

The algorithms for acoustic source localization based on PHAT filtering have been profusely used with good results in reverberant and noisy environments. However, there are very few studies that give a formal explanation of their robustness, most of them providing just an empirical validation or showing results on simulated data. In this work we present a novel analytical model for predicting the behavior of both the SRP-PHAT power maps and the GCC-PHAT functions. The results show that they are only affected by the signal bandwidth, the microphone array topology, and the room geometry, being independent of the spectral content of the received signal. The proposed model is shown to be valid in reverberant environments and under far and near field conditions. Using this result, an analysis study on how the aforementioned factors affect the SRP-PHAT power maps is presented providing well supported theoretical and practical considerations. The model validation is based on both synthetic and real data, obtaining in all cases a high accuracy of the model to reproduce the SRP-PHAT power maps, both in anechoic and non-anechoic scenarios, becoming thus an excellent tool to be exploited for the improvement of real world relevant applications related to acoustic localization.

1. Introduction

During the last three decades perceptual systems have been extensively studied. The objective is to automatically analyze complex information extracted from one or multiple sensors. These systems are the result of recent advances in sensor technologies, signal processing, and machine learning.

A fundamental task within this research area is the development of sensor technologies able to localize humans in indoor environments. While outdoor localization performance is very accurate, using GPS systems, indoor localization is still a largely unsolved problem. Localization of humans has a tremendous potential impact in diverse applied fields, opening new ways in how humans interact with machines.

In these scenarios, non-invasive technologies are preferred, so that no electronic or passive devices need to be carried by humans for localization. The two main non-invasive technologies for indoor localization are those based on cameras [1,2] and acoustic sensors [3,4]. Camera-based systems are very promising due to their reduced cost and the rich amount of information they provide. However camera-based systems still lack robust and fast methods that are resistant to varying light conditions and occlusions. Acoustic sensors give very rich information as humans naturally communicate with speech. Recent advances show that accurate localization with microphone arrays is possible and thus it is a promising technology in many applications.

* Corresponding author. Tel.: +34 918856918; fax: +34 918856591.
E-mail address: macias@depeca.uah.es (J. Macias-Guarasa).

http://dx.doi.org/10.1016/j.sigpro.2015.08.003
0165-1684/

Video and audio technologies are in fact very complementary in many ways and their fusion for localization has been extensively studied [5].

This paper focuses on audio-based localization in a very general indoor scenario with known geometry, where unknown wide-band audio sources are captured by a set of microphone arrays placed at known positions within the environment. From the various approaches in the literature we focus on systems based on computing the Steered Response Power (*SRP*) [6] of the signals captured in microphone arrays.

The gold standard for audio localization is based on combining SRP with the Phase Transform (PHAT) filtering, with varied modifications in order to provide improvements in precision and/or computational demands. SRP-PHAT has proved to be successful for localization in reverberant and noisy scenarios [4,7–9], but these proposals are only based on an empirical evaluation. Despite the success of the SRP-PHAT strategy, there are very few theoretical studies to understand its properties. To the best of our knowledge, there is no previous published work in which researchers provide an analytical model of the SRP-PHAT algorithm behavior, and none of the related published efforts attempt the validation of their findings using real data, thus limiting their application in real scenarios.

This paper's main contribution is to describe an analytical model of the SRP-PHAT power map that is valid in reverberant environments, under far and near field conditions, assuming reasonable constraints about sound propagation and its reflection on surfaces. We explicitly show that SRP-PHAT only depends on the signal bandwidth, the number of microphones, the geometry of the array and the geometry of the environment. As a consequence, the spectral content of the signal is not necessary to predict the SRP-PHAT maps. Using the proposed model we give well supported theoretical and practical insights about how each of these factors affect SRP-PHAT localization. Furthermore, since SRP can be understood as the sum of the generalized cross-correlations (GCC) of the chosen pairs of microphones [6], our model and conclusions can be easily extended to GCC-PHAT.

The model validation is carried out with both synthetic and real data, showing that our model is highly accurate to reproduce the SRP-PHAT power maps in reverberant scenarios. We believe that our model, given its parametric, analytical and differentiable nature, opens the path to interesting applications towards improving the performance of current localization systems, and automatic microphone arrays topology design, among others. Although this proposal is developed and evaluated for speech signals, we also believe that it is general enough to be easily extended to other wideband and narrowband acoustic signals.

1.1. Paper structure

The paper is structured as follows. In Section 2 we provide the state-of-the-art in acoustic source localization, focusing on the *SRP-PHAT* strategy. Section 3 describes the proposed model for the anechoic and non-anechoic cases. The experimental setup is detailed in Section 4 and model validation with synthetic and real data is addressed in Section 5. We study in Section 6 the effect of bandwidth and array topology in the model and finally, Section 7 summarizes the main conclusions and contributions of the paper and gives some ideas for future work.

2. State of the art

Acoustic source localization has received significant attention lately for people indoor localization, complementing other existing technologies, e.g., the CHIL (Computer in Human Interaction Loop) project [5]. The main motivation relies in the fact that acoustic source localization is essential in most of speech based human–machine interaction systems.

Existing approaches for acoustic source localization can be roughly divided into three categories [3,4,10]: (i) two-stage time delay, (ii) one-stage beamforming, and (iii) high-resolution spectral-estimation based methods. Methods in (iii) are not able to efficiently cope with real-world conditions (mainly noise and reverberation issues), making (i) and (ii) the leading methods.

Methods in (i) are composed of two stages: in the first step they estimate the time-difference of arrivals (TDOA) of signals between pairs of microphones [11]. This is usually done using generalized cross-correlation (GCC) techniques [12]. Among the possible weighting functions, the Phase Transform (PHAT) has been found to perform very well under realistic acoustical environments [13], leading to the GCC-PHAT [12] (also known as the Crosspower-Spectrum Phase [14]). There are also alternative methods, such as those based on Blind Source Separation [15], or those using a likelihood function of phase differences [16]. In a second step, the TDOA information is combined with knowledge of the microphones' positions, using optimization (maximum likelihood, least squares, spherical interpolation, etc.), to generate a spatial estimator of the source position [3,4,17]. The main problem with methods in (i) is their sensitivity to errors in the TDOA estimation, that can be hardly corrected if severe enough [6].

Beamforming [18] based techniques (ii) estimate the position of the source by sampling a set of possible spatial locations and computing a beamforming function at each location. The approach then chooses the source location that maximizes a statistic that is maximum when the target position matches the source location. For instance in SRP, which is the simplest beamforming method, the statistic is based on the signal power received when the microphone array is steered in the direction of a specific location. *SRP-PHAT* is a widely used algorithm for speaker localization based on beamforming. It was first proposed in [6] [1] and is a beamforming based method that combines the robustness of the steered beamforming methods with the insensitivity to signal conditions afforded by the PHAT filter. The classical *delay-and-sum* beamformer used in *SRP*

[1] Although the formulation is virtually identical to the *Global Coherence Field* (GCF) described in [14].

is replaced in *SRP-PHAT* by a *filter-and-sum* beamformer using PHAT filtering to weight the incoming signals.

The main problem with beamforming methods is their high computational cost, provided that they sample all potential positions of the space, labeling all local maxima as position candidates for acoustic sources.

SRP-PHAT is usually defined as a reference standard for source localization, because of its simplicity and robustness in reverberant and noisy environments, being a widely used algorithm for speaker localization [19–23].

The superior performance of SRP-PHAT is well supported by empirical evidence [4,8,9] yet there are just a few works aimed at giving formal explanations for its robustness. The authors of [24] evaluate SRP-PHAT and its variant β-PHAT so that they can emphasize the actual effect of the PHAT filtering, giving interesting insights about the effects of noise and reverberation in the localization performance. Unfortunately, the evaluation is purely experimental, without an analytic explanation, and based on simulated data. [13] shows that, under low noise and high reverberation conditions, SRP-PHAT is a special case of the maximum-likelihood estimator. Again, the results are based on simulated data and the assumptions made about the noise being gaussian do not hold when using real data [25]. The most recent known work in this area is described in [10]. Their approach is different to previous works, as they start from the signal models with some environmental assumptions, deriving an interesting analytic solution for the PHAT strategy. However, the formulation is only meant at explaining the PHAT robustness against reverberation, so that there is no attempt to further refine it by solving the frequency dependent terms, thus not allowing the derivation of further considerations to be used in practical applications. Additionally, the validation of their proposal is again based on simulated data, thus compromising its possible application to real-world scenarios.

In this paper, we derive an analytical model for predicting the behavior of the SRP-PHAT power maps, taking into account both the room geometry and the microphone array topology, and considering wideband signals. The proposal is valid for both near and far-field conditions, with a widely used signal model [26,10]. Our model also allows to intuitively analyze SRP-PHAT power maps, while being able to accurately represent their expected behavior. The model is validated using both synthetic and real data (as experiments in a real environment are essential for a model to be acceptable), and we provide a final discussion on some aspects of the model, easily extracted from the model formulation, thanks to its meaningful parametric analytical expression.

3. Generative steered response power model

3.1. Notation

Real scalar values are represented by lowercase letters (e.g., δ). Vectors are by default arranged column-wise and are represented by lowercase bold letters (e.g., $\mathbf{x}$). Uppercase letters are mostly reserved to define the size of vectors and sets (e.g., vector $\mathbf{x}=(x_1,\ldots,x_N)^\top$ is of size N). The l_2 norm of a vector $\|\mathbf{x}\|_2=\left(|x_1|^2+\cdots+|x_N|^2\right)^{1/2}$ will be written by default as $\|\cdot\|$ for simplicity. Calligraphic fonts are reserved to represent ranges or sets (e.g., $\mathbb{R}$ for real or generic sets $\mathcal{G}$). Continuous time signals are represented by scalar functions of t variable as for instance $x(t)$. Discrete time signals use k to denote discrete time samples. The Fourier transform of a continuous signal $x(t)$ is represented with complex function $X(\omega)$, with $X(\omega)^*$ being the complex-conjugate of $X(\omega)$ and $|X(\omega)|=\sqrt{X(\omega)^*X(\omega)}$. $\mathrm{Re}(X(\omega))$ and $\mathrm{Im}(X(\omega))$ are the real and imaginary parts of $X(\omega)$ respectively. We refer to $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ as the floor and ceiling rounding operators respectively.

3.2. Steered response power formulation

Let us assume we equip an indoor environment with an array of N_μ microphones $\mathcal{M}=\{\mathbf{m}_1,\mathbf{m}_2,\ldots,\mathbf{m}_{N_\mu}\}$, where $\mathbf{m}_n$ is a three-dimensional vector $\mathbf{m}_n=(m_{nx},m_{ny},m_{nz})^\top$ denoting the position of the microphone from a reference coordinate origin.

Given this setup, let us assume that an acoustic source is located at the generic position $\mathbf{r}=(r_x,r_y,r_z)^\top$, emitting an acoustic signal $x(t)$. We denote as $x_n(t)$ the signal received by a microphone located at $\mathbf{m}_n$, and as $\mathbf{q}=(q_x,q_y,q_z)^\top$ to a generic target location at which we steer the microphone array. We usually discretize the space using a finite set $\mathcal{Q}=\{\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_Q\}$ of Q three-dimensional vectors.

A *delay-and-sum* beamformer aligns the set of signals $x_1(t),\ldots,x_{N_\mu}(t)$, compensating the propagation delays from the target position $\mathbf{q}$ to each microphone $\mathbf{m}_n$. The resulting beamformed signal when the array is steered to $\mathbf{q}$ is defined as (also including the frequency domain expression):

$$y(t,\mathbf{q})=\sum_{n=1}^{N_\mu}x_n(t+\tau_n(\mathbf{q}))\overset{\mathcal{F}}{\leftrightarrow}Y(\omega,\mathbf{q})=\sum_{n=1}^{N_\mu}X_n(\omega)\mathrm{e}^{\mathrm{j}\omega\tau_n(\mathbf{q})},\tag{1}$$

where $X_n(\omega)$ is the Fourier Transform of $x_n(t)$, and $\tau_n(\mathbf{q})$ is the propagation delay between $\mathbf{m}_n$ and $\mathbf{q}$, which is calculated as $\tau_n(\mathbf{q})=\frac{1}{c}\|\mathbf{q}-\mathbf{m}_n\|$, where c is the speed of the sound in air.

The *filter-and-sum* beamformer is a generalization of the *delay-and-sum* beamformer, which applies adaptive filtering to the microphone signals. In this case, the signal received at microphone $\mathbf{m}_n$ is then filtered with $H_n(\omega)$. The beamformed signal when the array is steered to the position $\mathbf{q}$ is given in the frequency domain by

$$Y(\omega,\mathbf{q})=\sum_{n=1}^{N_\mu}H_n(\omega)X_n(\omega)\mathrm{e}^{\mathrm{j}\omega\tau_n(\mathbf{q})}.\tag{2}$$

The *Steered Response Power* (SRP) can be expressed as the output power of the signal received from a *filter-and-sum* beamformer of N_μ elements [4]:

$$\begin{aligned}P(\mathbf{q})&=\int_{-\infty}^{-\infty}|Y(\omega,\mathbf{q})|^2\,\mathrm{d}\omega\\&=\sum_{i=1}^{N_\mu}\sum_{j=1}^{N_\mu}\int_{-\omega_0}^{\omega_0}H_i(\omega)H_j^*(\omega)X_i(\omega)X_j^*(\omega)\mathrm{e}^{\mathrm{j}\omega\tau_i(\mathbf{q})}\mathrm{e}^{-\mathrm{j}\omega\tau_j(\mathbf{q})}\,\mathrm{d}\omega,\end{aligned}\tag{3}$$

where $P(\mathbf{q})$ is the power received at position $\mathbf{q}$, and $x_n(t)$ is a baseband signal with bandwidth ω_0 ($X_n(\omega)=0,\forall\omega>\omega_0$).

3.3. Steered response power with PHAT

The PHAT filter has been typically applied in the context of the Generalized Cross Correlation (GCC) [12]. The GCC of two signals $X_i(\omega)$ and $X_j(\omega)$ is defined as

$$R_{i,j}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Psi_{ij}(\omega) X_i(\omega) X_j^*(\omega) \mathrm{e}^{-\mathrm{j}\omega\tau}\, \mathrm{d}\omega, \tag{4}$$

where $\Psi_{ij}(\omega)$ is a weighting function that is chosen to optimize a performance criteria. Apart from PHAT, many other weighting functions have been proposed, such as the Smoothed Coherence Transform (SCOT) [27]. In PHAT, weighting $\Psi_{ij}(\Omega)$ is chosen as

$$\Psi_{ij}(\omega) = \frac{1}{|X_i(\omega)X_j^*(\omega)|} = \frac{1}{|X_i(\omega)|\,|X_j(\omega)|}, \tag{5}$$

where we also assume that $|X_i(\omega)|$ and $|X_j(\omega)|$ must be strictly greater than 0 within the signal bandwidth. PHAT weighting empirically reduces the influence of multipath in the GCC of a signal arriving to two different microphones in a reverberant acoustic environment.

Applying the PHAT filter as defined in Eq. (5) to the received power in Eq. (3), is equivalent to using, on each microphone, a filter whose transfer function is $H_n(\omega) = |X_n(\omega)|^{-1}$:

$$\begin{aligned} P(\mathbf{q}) &= 2\pi \sum_{i=1}^{N_\mu}\sum_{j=1}^{N_\mu} R_{ij}(\Delta\tau_{ij}(\mathbf{q})) \\ &= \sum_{i=1}^{N_\mu}\sum_{j=1}^{N_\mu} \int_{-\omega_0}^{\omega_0} \frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)|\,|X_j(\omega)|} \mathrm{e}^{\mathrm{j}\omega\Delta\tau_{ij}(\mathbf{q})}\, \mathrm{d}\omega \\ &= 2N_\mu\omega_0 + \sum_{i=1}^{N_\mu}\sum_{\substack{j=1\\ j\neq i}}^{N_\mu} \int_{-\omega_0}^{\omega_0} \frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)|\,|X_j(\omega)|} \mathrm{e}^{\mathrm{j}\omega\Delta\tau_{ij}(\mathbf{q})}\, \mathrm{d}\omega, \end{aligned} \tag{6}$$

where $\Delta\tau_{ij}(\mathbf{q}) = (\tau_i(\mathbf{q}) - \tau_j(\mathbf{q}))$, and where we have explicitly stated the calculation of the SRP as a function of the generalized cross correlations [6]. In this expression we have also considered that for $i=j$, the integral represents the power received by each single microphone after applying PHAT. These terms have a trivial solution $\left(\int_{-\omega_0}^{\omega_0} \frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)|\,|X_j(\omega)|}\,\mathrm{d}\omega = 2\omega_0, \forall i=j\right)$ which does not depend on the steering position, so they only represent a known offset ($2N_\mu\omega_0$) that can be easily removed from Eq. (6). Without loss of generality, we will not take this offset term into account hereinafter (this will imply the appearance of negative values in $P(\mathbf{q})$).

We group the microphones in different pairs, described as elements in a set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_P}\}$, where $\mathbf{p}_j = \{\mathbf{m}_{j_1}, \mathbf{m}_{j_2}\}$ is composed of two three-dimensional vectors, $\mathbf{m}_{j_1} \in \mathcal{M}$ and $\mathbf{m}_{j_2} \in \mathcal{M}$, with $\mathbf{m}_{j_1} \neq \mathbf{m}_{j_2}$, describing the spatial location of the microphones in pair j. If all microphone pairs are allowed then $N_P = N_\mu(N_\mu - 1)/2$.

Eq. (6) can be rewritten taking into account the contributions of each pair of microphones:

$$\begin{aligned} P(\mathbf{q}) &= \sum_{j=1}^{N_P} \int_{-\omega_0}^{\omega_0} \frac{X_{j_1}(\omega)X_{j_2}^*(\omega)\mathrm{e}^{\mathrm{j}\omega\Delta\tau(\mathbf{p}_j,\mathbf{q})} + X_{j_1}^*(\omega)X_{j_2}(\omega)\mathrm{e}^{-\mathrm{j}\omega\Delta\tau(\mathbf{p}_j,\mathbf{q})}}{|X_{j_1}(\omega)|\,|X_{j_2}(\omega)|}\,\mathrm{d}\omega \\ &= \sum_{j=1}^{N_P}\left[\int_{-\omega_0}^{\omega_0} 2\frac{\mathrm{Re}\left(X_{j_1}(\omega)X_{j_2}^*(\omega)\mathrm{e}^{\mathrm{j}\omega\Delta\tau(\mathbf{p}_j,\mathbf{q})}\right)}{|X_{j_1}(\omega)|\,|X_{j_2}(\omega)|}\,\mathrm{d}\omega\right] \\ &= 4\pi\sum_{j=1}^{N_P} R_{j_1 j_2}\left(\Delta\tau(\mathbf{p}_j,\mathbf{q})\right), \end{aligned} \tag{7}$$

where $\Delta\tau(\mathbf{p}_j,\mathbf{q}) = (\tau_{j_1}(\mathbf{q}) - \tau_{j_2}(\mathbf{q}))$ is the difference in arrival times of the acoustic signal to reach the microphones in pair $\mathbf{p}_j$ ($\mathbf{m}_{j_1}$ and $\mathbf{m}_{j_2}$), that is, the required delay to steer the microphone pair $\mathbf{p}_j$ to the location $\mathbf{q}$. The last expression in (7) will allow us to extend the conclusions of the paper to both SRP-PHAT and GCC-PHAT, as it states the calculation of the SRP-PHAT as the sum of the GCC-PHAT functions $R_{j_1 j_2}\left(\Delta\tau(\mathbf{p}_j,\mathbf{q})\right)$, for the considered N_p pairs of microphones [6].

3.4. Anechoic propagation

In the simplest scenario, when anechoic propagation is assumed, the signal received by each microphone is a delayed and attenuated version of the acoustic source signal $x(t)$:

$$x_n(t) = \alpha_n x(t - \tau_n(\mathbf{r})) \overset{\mathcal{F}}{\leftrightarrow} X_n(\omega) = \alpha_n X(\omega)\mathrm{e}^{-\mathrm{j}\omega\tau_n(\mathbf{r})}, \tag{8}$$

where $\alpha_n = \frac{1}{4\pi c\tau_n(\mathbf{r})}$ is a distance-related attenuation assuming spherical propagation [28].

In this case, Eq. (7) can be simplified as follows:

$$\begin{aligned} P(\mathbf{q}) &= \sum_{j=1}^{N_P}\int_{-\omega_0}^{\omega_0} 2\frac{\mathrm{Re}\left(\alpha_{j_1}\,\alpha_{j_2}|X(\omega)|^2\mathrm{e}^{\mathrm{j}\omega\left(\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau(\mathbf{p}_j,\mathbf{r})\right)}\right)}{\alpha_{j_1}\alpha_{j_2}|X(\omega)|^2}\,\mathrm{d}\omega \\ &= \sum_{j=1}^{N_P}\int_{-\omega_0}^{\omega_0} 2\cos\left(\omega(\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)\mathrm{d}\omega \\ &= \sum_{j=1}^{N_P}\left[\frac{4\sin\left(\omega_0(\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)}{\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau(\mathbf{p}_j,\mathbf{r})}\right], \end{aligned} \tag{9}$$

where α_{j_1} and α_{j_2} are the attenuation factors for the first and second microphones of the $\mathbf{p}_j$ pair respectively. Each of the terms in the sum of Eq. (9) corresponds to the GCC-PHAT model for each microphone pair.

From (9) we derive the following result for anechoic propagation:

In anechoic conditions the steered response pattern with PHAT filtering does not depend on the spectral content of the measured signal. It is only affected by the signal bandwidth, the number of microphones and the distance between them.

The signal independence we show here is one of the reasons why the PHAT approach performs so well in real scenarios with human speech involved, in which the signal content is largely unknown.

3.5. Non-anechoic propagation

The anechoic propagation model described in Eq. (8) only considers the direct path between the source and each microphone. In real indoor scenarios, where many reflective surfaces (i.e. walls, ceiling, floor, etc.) are present, some acoustic energy from the source is reflected on these surfaces and reaches the microphones.

The room's effect on the signal can be generically characterized by the room impulse response $h(t,\mathbf{r},\mathbf{m}_n)$ [29], which depends on the position of both source and receptor (microphone). The signal received in each microphone can be calculated as follows:

$$x_n(t) = h(t,\mathbf{r},\mathbf{m}_n) * x(t) + n(t), \tag{10}$$

where $*$ is the convolution operator and $n(t)$ represents additive noise.

In [28] a very detailed signal propagation model for a general enclosure is presented. However, it is too complex to be handled, so that certain (reasonable) assumptions are usually employed in the literature:

- The sources emit spherical sound waves. We know that in real situations the human head shows a more complex, frequency dependent radiation pattern [30].
- The medium is homogeneous and non-dispersive (so that the sound speed c is constant and frequency independent), and lossless (it does not absorb energy from propagating waves).
- The acoustic signal travels from the source position to the destination microphone through different paths (each one composed of several subpaths), as the signal is reflected by the corresponding surfaces.
- The frequency shaping and dependence effect of the source, microphones and surface reflections are ignored, assuming them to be constant and frequency independent for all signal paths.
- The background noise is assumed negligible [10] and not correlated with the reverberation effect [13], as we are mainly concerned with the reverberation effects .

The aforementioned restrictions are equivalent to consider the room impulse response as the sum of a finite number of delayed Dirac deltas, each of them multiplied by a factor inversely proportional to the delay:

$$h(t,\mathbf{r},\mathbf{m}_n) = \sum_{k=0}^{K} \frac{A_k \delta(t-\tau_{n,k}(\mathbf{r}))}{4\pi c \tau_{n,k}(\mathbf{r})}, \tag{11}$$

where $K+1$ is the number of different paths from the source $\mathbf{r}$ to the destination microphone $\mathbf{m}_n$, and $\tau_{n,k}(\mathbf{r})$ is the propagation delay between $\mathbf{r}$ and $\mathbf{m}_n$ for path k. A_k is the attenuation factor due to reflections along the path k. The index $k=0$ will be used for referring the direct path, therefore $\tau_{n,0}(\mathbf{r}) = \frac{1}{c}\|\mathbf{r}-\mathbf{m}_n\|$ and $A_0 = 1$.

In conclusion, applying the former considerations to Eq. (10), we obtain the following expression for the signal received in each microphone:

$$x_n(t) = \sum_{k=0}^{K} \frac{A_k x(t-\tau_{n,k}(\mathbf{r}))}{4\pi c \tau_{n,k}(\mathbf{r})} = \sum_{k=0}^{K} \frac{G}{\tau^*_{n,k}(\mathbf{r})} x\big(t-\tau_{n,k}(\mathbf{r})\big), \tag{12}$$

where $G = \frac{1}{4\pi c}$ is constant among all paths and $\tau^*_{n,k}(\mathbf{r}) = \frac{\tau_{n,k}(\mathbf{r})}{A_k}$ combines the effects on the amplitude of A_k and $\tau_{n,k}(\mathbf{r})$.

The Fourier Transform of Eq. (12) is

$$X_n(\omega) = \sum_{k=0}^{K} \frac{G}{\tau^*_{n,k}(\mathbf{r})} X(\omega) \mathrm{e}^{-\mathrm{j}\tau_{n,k}(\mathbf{r})}. \tag{13}$$

Using the signal model described by Eq. (13) in Eq. (7), we get

$$\begin{aligned} P(\mathbf{q}) &= \sum_{j=1}^{N_P} \int_{-\omega_0}^{\omega_0} 2 \frac{\mathrm{Re}\left(\left(\sum_{k=0}^{K} \frac{GX(\omega)}{\tau^*_{j_1 k}} \mathrm{e}^{-\mathrm{j}\omega\tau_{j_1 k}}\right)\left(\sum_{l=0}^{K} \frac{GX^*(\omega)}{\tau^*_{j_2 l}} \mathrm{e}^{+\mathrm{j}\omega\tau_{j_2 l}}\right)\mathrm{e}^{\mathrm{j}\omega\Delta\tau(\mathbf{p}_j,\mathbf{q})}\right)}{\left|\sum_{k=0}^{K} \frac{GX(\omega)}{\tau^*_{j_1 k}} \mathrm{e}^{-\mathrm{j}\omega\tau_{j_1 k}}\right|\left|\sum_{l=0}^{K} \frac{GX(\omega)}{\tau^*_{j_2 l}} \mathrm{e}^{-\mathrm{j}\omega\tau_{j_2 l}}\right|} \mathrm{d}\omega \\ &= \sum_{j=1}^{N_P} \int_{-\omega_0}^{\omega_0} 2 \frac{\mathrm{Re}\left(\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{\mathrm{e}^{\mathrm{j}\omega(\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r}))}}{\tau^*_{j_1 k}\tau^*_{j_2 l}}\right)}{\left|\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_1 k}+\tau_{j_2 l})}}{\tau^*_{j_1 k}\tau^*_{j_2 l}}\right|} \mathrm{d}\omega \\ &= \sum_{j=1}^{N_P}\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{2}{\tau^*_{j_1 k}\tau^*_{j_2 l}} \int_{-\omega_0}^{\omega_0} \frac{\cos\left(\omega(\Delta\tau(\mathbf{p}_j,\mathbf{q}) - \Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r}))\right)}{\left|\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_1 k}+\tau_{j_2 l})}}{\tau^*_{j_1 k}\tau^*_{j_2 l}}\right|} \mathrm{d}\omega \end{aligned} \tag{14}$$

where, in order to simplify the notation, $\tau_{n,k}$ and $\tau^*_{n,k}$ are used instead of $\tau_{n,k}(\mathbf{r})$ and $\tau^*_{n,k}(\mathbf{r})$. Indices j_1 and j_2 refer to the first and second microphones of the pair $\mathbf{p}_j$, respectively. Therefore, $\Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r}) = \tau_{j_1}(\mathbf{r};k) - \tau_{j_2}(\mathbf{r};k)$ is the difference in arrival times of the acoustic signal to reach the first microphone of the pair $\mathbf{p}_j$ ($\mathbf{m}_{j1}$) using path k, and the second microphone of the pair $\mathbf{p}_j$ ($\mathbf{m}_{j2}$) using path l.

The denominator term in Eq. (14) can be developed as follows:

$$\begin{aligned} &\left|\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_1 k}+\tau_{j_2 l})}}{\tau^*_{j_1 k}\tau^*_{j_2 l}}\right| \\ &= \left(\sum_{k=0}^{K}\sum_{l=0}^{K}\sum_{m=0}^{K}\sum_{n=0}^{K} \frac{\cos\left(\omega(\tau_{j_1 k}+\tau_{j_2 l}-\tau_{j_1 m}-\tau_{j_2 n})\right)}{\tau^*_{j_1 k}\tau^*_{j_2 l}\tau^*_{j_1 m}\tau^*_{j_2 n}}\right)^{1/2}. \end{aligned} \tag{15}$$

The last expression in Eq. (15) can be rewritten as

$$\left(\sum_{k=0}^{K}\sum_{l=0}^{K} \frac{1}{(\tau^*_{j_1 k}\tau^*_{j_2 l})^2} + \sum_{k_1}\sum_{l_1}\sum_{k_2}\sum_{l_2} \frac{\cos\left(\omega(\tau_{j_1 k_1}+\tau_{j_2 l_1}-\tau_{j_1 k_2}-\tau_{j_2 l_2})\right)}{\tau^*_{j_1 k_1}\tau^*_{j_2 l_1}\tau^*_{j_1 k_2}\tau^*_{j_2 l_2}}\right)^{1/2}, \tag{16}$$

where the first term includes all the terms in Eq. (15) such that $k=m$ and $l=n$.

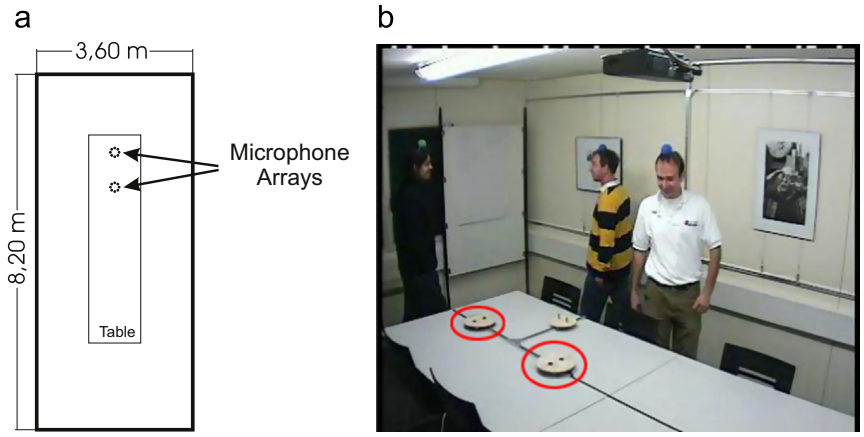


Fig. 1. IDIAP Smart Meeting Room for AV16.3 recordings. (a) Room layout showing the centered table, and the microphones arranged in two circular arrays. (b) Sample of recorded video frame showing the arrays area. (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

Based on the Law of large numbers, and under the reasonable assumption that the distribution of the cosine argument in Eq. (16) is uniform, it can be demonstrated that the expected value of the second term in Eq. (16) is zero [10, Chapter 2.3]. Consequently, we can approximate the denominator in Eq. (14) as $D(\mathbf{r})=\sqrt{\sum_{k=0}^{K}\sum_{l=0}^{K}\frac{1}{(\tau^*_{j_1k}\tau^*_{j_2l})^2}}$, which is a frequency-independent term.

From the discussion above, we finally get the expression of the SRP at any given location $\mathbf{q}$:

$$\begin{aligned}P(\mathbf{q})&=\sum_{j=1}^{N_P}\sum_{k=0}^{K}\sum_{l=0}^{K}\frac{2}{\tau^*_{j_1k}\tau^*_{j_2l}}\int_{-\omega_0}^{\omega_0}\frac{\cos\left(\omega(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r}))\right)}{D(\mathbf{r})}\,\mathrm{d}\omega\\&=\sum_{j=1}^{N_P}\sum_{k=0}^{K}\sum_{l=0}^{K}\frac{4}{\tau^*_{j_1k}\tau^*_{j_2l}D(\mathbf{r})}\frac{\sin\left(\omega_0(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r}))\right)}{\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r})}\\&=\sum_{j=1}^{N_P}\left[\sum_{k=0}^{K}\sum_{l=0}^{K}\frac{4\omega_0}{\tau^*_{j_1k}\tau^*_{j_2l}D(\mathbf{r})}\mathrm{sinc}(\omega_0(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau_{k,l}(\mathbf{p}_j,\mathbf{r})))\right],\end{aligned}\tag{17}$$

which is a *linear combination of sinc functions* ($\mathrm{sinc}(x)=\frac{\sin(x)}{x}$). Again, each of the terms in the sum for j of Eq. (17) (the term within brackets) corresponds to the GCC-PHAT model for each microphone pair. Note that the anechoic pattern expressed in Eq. (9) is a particular case of the non-anechoic pattern when $K=0$.

From (17) we derive the following result for reverberant conditions:

In non-anechoic conditions, the steered response pattern with PHAT filtering does not depend on the spectral content of the measured signal. It is only affected by the signal bandwidth, the number of microphones, the distance between them and the environment's impulse response.

3.6. Practical considerations

In the previous section, continuous time signals were assumed and the continuous time Fourier Transform was used to determine the spectral content of the acoustic signals. In practice we use discrete representations of the signals and we apply the Discrete Fourier Transform (DFT) to a given signal window.

Considering that $T_s=\frac{1}{f_s}$ is the sampling period, $\hat{y}(n,\mathbf{q})=w(n)y(nT_s,\mathbf{q})$ is the windowed and sampled version of $y(t,\mathbf{q})$ defined in Eq. (1), where $w(n)$ is the used window function.

Assuming that we are working with a signal window of N_W samples, we can write

$$P_{DFT}(\mathbf{q})=\sum_{n=0}^{N_W-1}|\hat{y}(n,\mathbf{q})|^2=\frac{1}{N_F}\sum_{k=0}^{N_F-1}|\hat{Y}(k,\mathbf{q})|^2,\tag{18}$$

in which we have applied Parseval's Theorem, and where $\hat{Y}(k,\mathbf{q})$ is the DFT of $\hat{y}(n,\mathbf{q})$[2].

Given a properly selected windowing function, we can approximate $\hat{Y}(k,\mathbf{q})\approx Y\left(\frac{2\pi k}{N_FT_s},\mathbf{q}\right)$, to avoid further complicating the mathematical development. Therefore

$$\begin{aligned}P_{DFT}(\mathbf{q})&\approx\frac{1}{N_F}\sum_{k=0}^{N_F-1}\left|Y\left(\frac{2\pi k}{N_FT_s},\mathbf{q}\right)\right|^2\\&=\frac{1}{N_F}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=0}^{N_F-1}\frac{X_i\left(\frac{2\pi k}{N_FT_s}\right)X_j^*\left(\frac{2\pi k}{N_FT_s}\right)}{\left|X_i\left(\frac{2\pi k}{N_FT_s}\right)\right|\left|X_j\left(\frac{2\pi k}{N_FT_s}\right)\right|}\mathrm{e}^{\mathrm{j}(2\pi k/N_FT_s)\Delta\tau_{ij}(\mathbf{q})}.\end{aligned}\tag{19}$$

Since the DFT is periodic in frequency, $X(\omega)$ is bandlimited, and $f_0<\frac{f_s}{2}$, we can restrict the sum to $k\in[-M_0,M_0]$, being M_0 the DFT index corresponding to the signal bandwidth (ω_0):

$$M_0=\left\lfloor\frac{\omega_0T_sN_F}{2\pi}\right\rfloor=\left\lfloor\frac{f_0N_F}{f_s}\right\rfloor.\tag{20}$$

Grouping in pairs, in the same way as we did in Section 3.3, Eq. (19) can be rewritten as

$$P_{DFT}(\mathbf{q})=\frac{1}{N_F}\sum_{j=1}^{N_P}\sum_{k=-M_0}^{M_0}2\frac{\mathrm{Re}\left(X_{j_1}\left(\frac{2\pi k}{N_FT_s}\right)X_{j_2}^*\left(\frac{2\pi k}{N_FT_s}\right)\mathrm{e}^{\mathrm{j}(2\pi k/N_FT_s)\Delta\tau(\mathbf{p}_j,\mathbf{q})}\right)}{\left|X_{j_1}\left(\frac{2\pi k}{N_FT_s}\right)\right|\left|X_{j_2}\left(\frac{2\pi k}{N_FT_s}\right)\right|}.\tag{21}$$

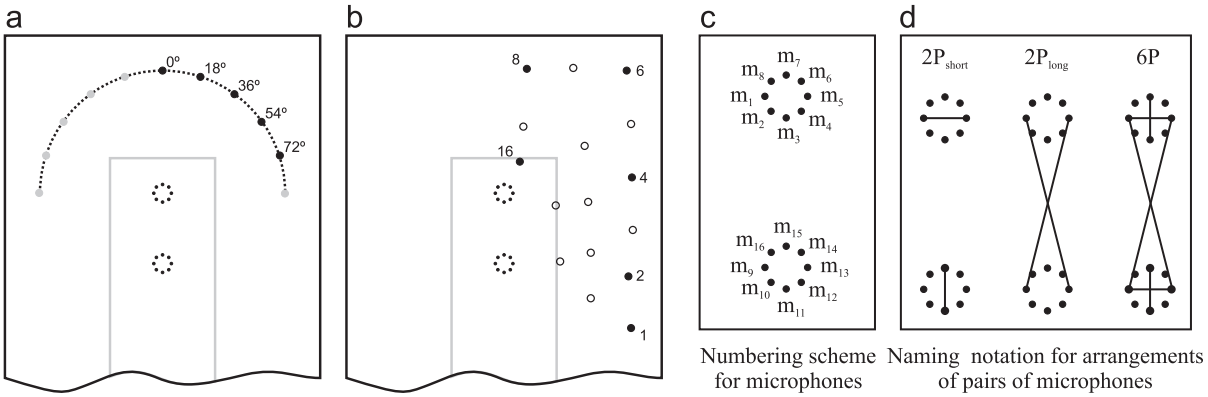


Fig. 2. Geometrical details for the experiments carried out. In (a) and (b) only the relevant room section is shown. (a) Positions for validation with simulated data. (b) Positions for validation with real data. (c) Numbering scheme for microphones and naming notation for arrangements of pairs of microphones.

Taking into account the propagation model described in (8), we can rewrite (21) as

$$P_{DFT}(\mathbf{q}) = \frac{1}{N_F}\sum_{j=1}^{N_P}\sum_{k=-M_0}^{M_0} 2\cos\left(\frac{2\pi k}{N_F T_s}(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau(\mathbf{p}_j,\mathbf{r}))\right) = \frac{1}{N_F}\sum_{j=1}^{N_P}\frac{2\sin\left(\frac{2\pi(M_0-0.5)}{N_F T_s}(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)}{\sin\left(\frac{\pi}{N_F T_s}(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)}, \quad (22)$$

in which the DFT effects are implicit. Note that the non-anechoic model described in Eq. (12) can be also applied, resulting in

$$P_{DFT}(\mathbf{q}) = \frac{1}{N_F}\sum_{j=1}^{N_P}\left[\sum_{k=0}^{K}\sum_{l=0}^{K}\frac{2\sin\left(\frac{2\pi(M_0-0.5)}{N_F T_s}(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)}{\tau^*_{j_1 k}\tau^*_{j_2 l}D(\mathbf{r})\sin\left(\frac{\pi}{N_F T_s}(\Delta\tau(\mathbf{p}_j,\mathbf{q})-\Delta\tau(\mathbf{p}_j,\mathbf{r}))\right)}\right], \quad (23)$$

where the term within brackets corresponds to the GCC-PHAT model for each microphone pair.

In practical applications, the *Fast Fourier Transform* (FFT) is used for the DFT, and in order to avoid its circular convolution effects, zero-padding is needed. So, assuming that we are working with a signal window duration of T_W seconds, sampled at $f_s = \frac{1}{T_s}$ Hz, the number of samples needed after zero-padding is at least $N_F = 2T_W f_s$. In addition, for efficient execution of the FFT, N_F must be a power of 2, so that the number of samples needed is $N_F = 2^{\{\lceil \log_2(T_W f_s)\rceil + 1\}}$.

If N_F is high enough (more precise details about this approximation can be found in Section 6.3), it can be easily shown that Eqs. (22) and (9) are proportional, so that $P_{DFT}(\mathbf{q}) = \frac{1}{2\pi f_s}P(\mathbf{q})$, thus validating also $P_{DFT}(\mathbf{q})$ to accurately represent $P(\mathbf{q})$.

4. Experimental setup

The model proposed here has been evaluated using audio recordings from the AV16.3 database [31], an audio–visual corpus recorded in the *Smart Meeting Room* of the IDIAP research institute, in Switzerland.

The IDIAP Meeting Room (shown in Fig. 1) is a 8.2 m × 3.6 m × 2.4 m rectangular space containing a centrally located 4.8 m × 1.2 m rectangular table, on top of which two circular microphone arrays of 10 cm radius are located, each of them composed by 8 microphones. The centers of the two arrays are separated by 80 cm and the origin of coordinates is located in the middle point between the two arrays. The arrays can be also seen in Fig. 2a and b, in which each one shows different scenarios that were used in the experiments, displaying only the relevant section of the room. Fig. 2c shows the microphone numbering notation, and different microphone arrangements that were tested: $2P_{short}$ and $2P_{long}$ using two microphone pairs (with different microphone separations), and $6P$ using six microphone pairs. Experiments considering *all* microphone pairs have also been done. A detailed description of the meeting room can be found in [32].

The audio recordings are synchronously sampled at 16 KHz, and the complete database along with the corresponding annotation files containing the recordings ground truth is fully accessible on-line at [33]. It is composed by several sequences or recordings which range in the number of speakers involved and their activity. In this paper we will just focus on sequence 01, in which a single male speaker generates digit strings in 16 static positions (which can be seen as small circles in Fig. 2b), distributed along the room. The sequence duration accounts for 208 s in total, with 2248 ground truth frames.

The audio sequence is assigned a corresponding annotation file containing the ground truth positions (3D coordinates) of the speaker's mouth at every time frame in which that speaker was talking. The frame shift resolution was defined to be 40 ms.

5. Model validation

In this section the validity of the proposed model is checked, first against simulated data, and then using real data. In both cases the results will show that the model fits the validation data with high precision. The validation strategy is designed as follows: First we assess the validity

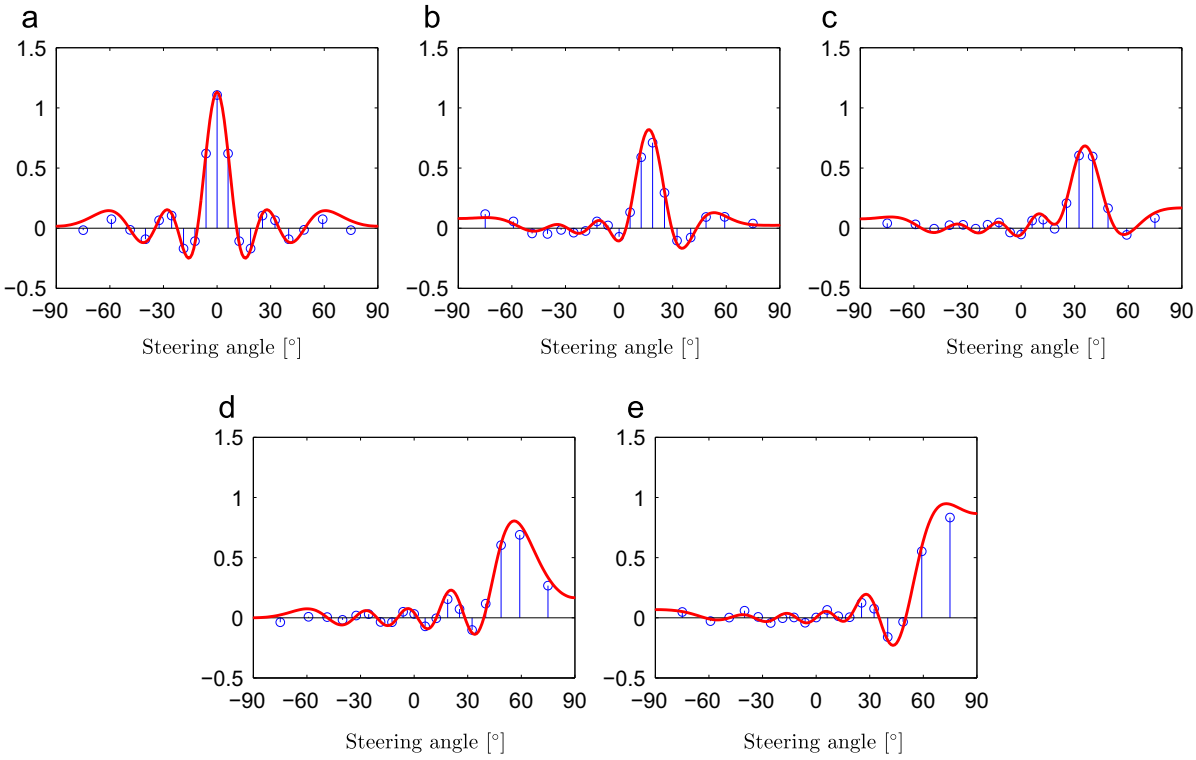


Fig. 3. Comparison between the steered power response generated by the model (solid line) and that calculated using simulated waveforms in the AV16.3 environment (stems). Results for the speaker in given angles and the array steered from -90° to $+90^\circ$ are shown. (a) 0°, (b) 18°, (c) 36°, (d) 54°, (e) 72°.

of the model to be able to accurately predict the contribution of a single microphone pair to the SRP-PHAT map (each GCC-PHAT model term in Eqs. (9), (17) and (23)), using simulated data: if the accuracy is high, the model is also expected to be useful to predict the SRP-PHAT behavior, given that the SRP-PHAT map is built by adding the contribution of all the selected microphone pairs, as expressed in Eq. (7). Next, and also on simulated data, we show how the modeled single pair contributions are combined to form the full SRP-PHAT maps, showing the accuracy of the model prediction. Finally, we address the evaluation using real data, checking the ability of the model to accurately predict the real full SRP-PHAT maps for varying geometrical conditions.

5.1. *Validation with simulated data: GCC-PHAT prediction performance*

The first validation experiments are based on simulated data, with the following characteristics (refer to Fig. 2 for geometrical details):

- The sound propagation effects have been simulated considering a room with the same dimensions as the IDIAP Meeting Room. The simulated waveforms have been generated using the impulse responses from the source to each microphone, computed using the image method model [34]. To ease the simulation and interpretation, only first order reflections have been considered, with all reflection coefficients set to 0.95 (corresponding to an absorption coefficient of 0.10).
- A single microphone pair is used, with a separation of 20 cm between them, corresponding to microphones $\mathbf{m}_1$ and $\mathbf{m}_5$ in the AV16.3 setup.
- The received array power is calculated steering the array towards different angles along a semi-circle centered at the microphone pair center, with radius of 1.4 m.
- The speaker is moving also along the same semi-circle.
- In order to perform an angular sweep, 181 different angular positions separated 1° between each other were evaluated.

The validation was done on the 181 different source positions, and some example results are shown in Fig. 3, in which we have evaluated the situation for the speaker located at different angles (0°, 18°, 36°, 54° and 72°, as shown in Fig. 2a), while the array was being steered from -90 to $+90$. Fig. 3 shows, for each steering angle, the comparison between the steered response power generated by the model (shown as solid lines and corresponding to the GCC-PHAT model for the given microphone pair (the term within brackets in Eq. (23))) with the actual steered power calculated using the simulated waveforms as they propagate in the environment (shown as *stems*, corresponding to the GCC-PHAT function for the given microphone pair). Note that the resolution of the simulated measurements is restricted by the sample rate: the angular position of the measurements correspond with the samples of the

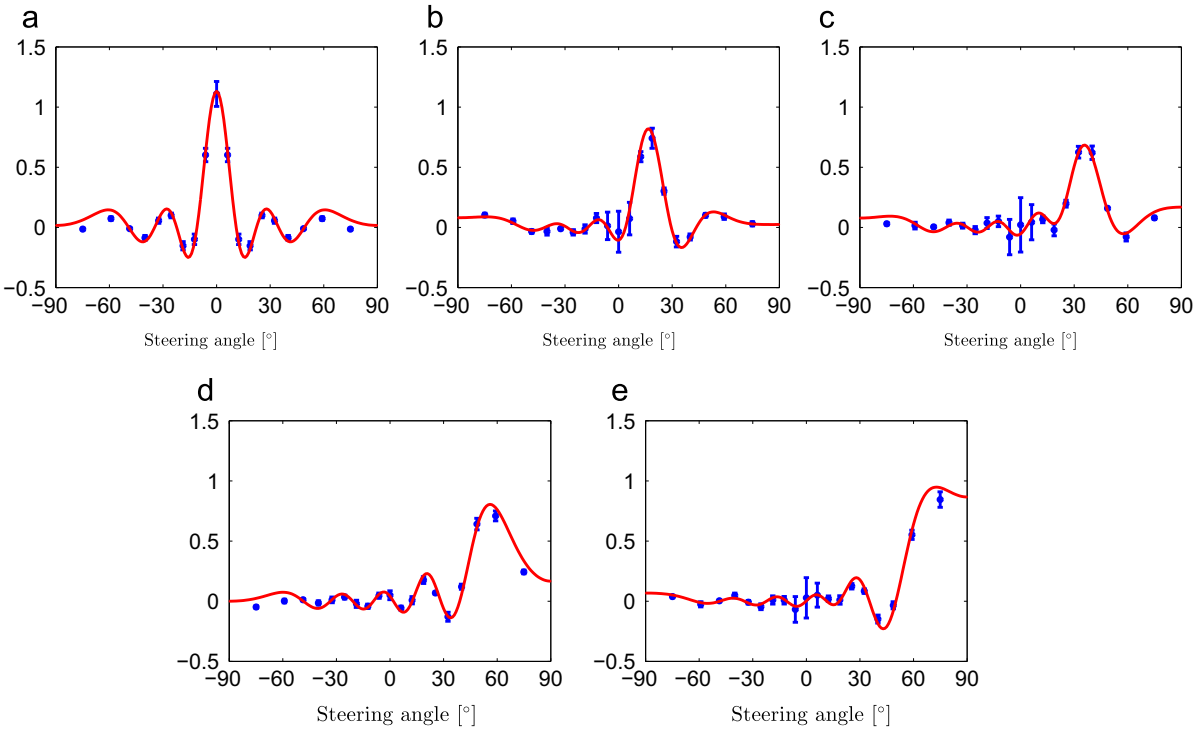


Fig. 4. Comparison between the steered power response generated by the model (solid line) and the average behaviour calculated using simulated waveforms in the AV16.3 environment, plus the standard deviation of the error (between the model and the simulation) for each angle (intervals). Results for different angles are shown. (a) 0°, (b) 18°, (c) 36°, (d) 54°, (e) 72°.

correlation function. On the other hand, the proposed model is continuous, what means that it could be used for increasing the resolution in localization.

To further assess the model precision, in Fig. 4, we show the comparison between the steered response power generated by the model for each steering angle (shown as a solid line) and the average actual steered power calculated using the simulated waveforms (shown as *dots*). In Fig. 4 we have also added the intervals corresponding to the standard deviation of the error between the model and the simulated response. It can be seen that the error deviation is very small along the angular variation, with a slight increment around the 0° angle, due to the increased ambiguity given the symmetry of the room, that leads to different acoustic paths to generate the same delays.

To summarize, from Figs. 3 and 4, we can clearly see that the model is able to reproduce the main correlation lobes generated by the effect of the direct path with high precision, and also the spurious effects caused by reverberation (shown by lower amplitude *sinc*-like functions), so that we can conclude that the agreement between the model predictions and the simulated data is very high, thus validating it from a theoretical perspective.

5.2. Validation with simulated data: SRP-PHAT prediction performance

The second validation experiments are also based on simulated data, following the same characteristics described in the previous section. In this case we are interested in showing how the proposed model is able to accurately predict simulated SRP-PHAT maps, by combining the GCC-PHAT responses of several microphone pairs.

For the IDIAP meeting room, all reflection coefficients have been set to 0.86 (given that we do not have the precise details on the materials of the real room surfaces, we have estimated this reflection coefficient making reasonable assumptions on these materials (painted concrete walls, carpet on the floor, and fiberboard in the ceiling), averaging their frequency absorption coefficients using the data available at [35], and taking also into account the area of the corresponding surfaces (using Sabine's formula [36]). The estimated RT_{60} coefficient for this setup is 392 ms.

In the graphics showing the comparison of the SRP-PHAT power maps (predicted by Eq. (23), or calculated), the plots are provided from a top view of the room, spanning the full room plan for a regular two-dimensional grid of 10 cm, at a height of 61 cm above the microphone arrays (this height was the ground truth one for sequence 01). The green point represents the real (*ground truth*) speaker position, and the black dots represent the positions of the microphones used. In all the comparisons, three graphics are plotted:

- The graphics on the left will show the SRP-PHAT acoustic power maps generated by the proposed model (for example, the left graphic in Fig. 5a).
- The graphics in the middle will show the SRP-PHAT acoustic power maps calculated using the acoustic waveforms for a single selected frame (for example, the middle graphic in Fig. 5a).
- The graphics on the right will show the average SRP-PHAT acoustic power maps, averaging for all the

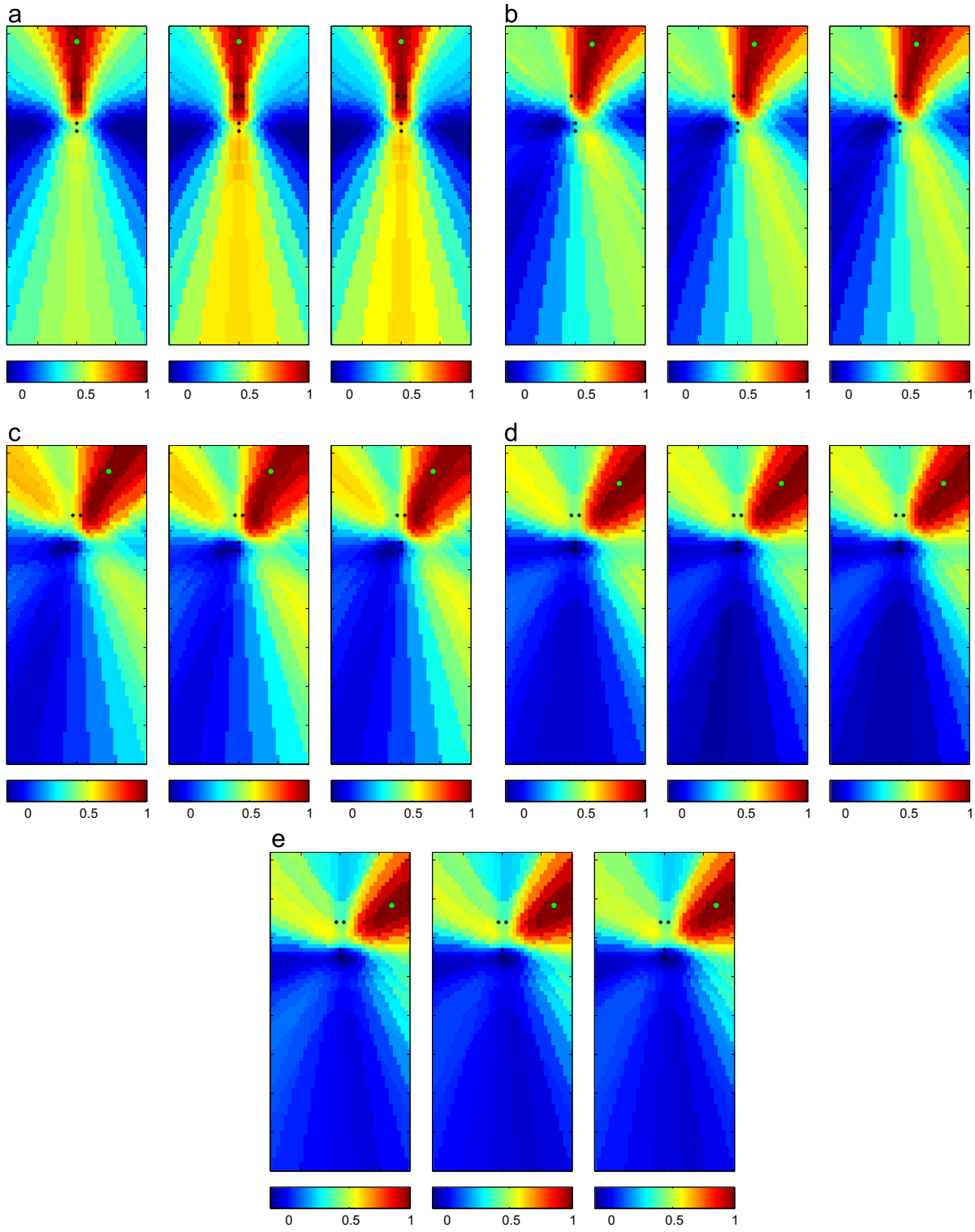


Fig. 5. Comparison between the SRP-PHAT map predicted by the model (left graphics), the simulated SRP-PHAT map (middle graphics), and the average (simulated) SRP-PHAT map (right graphics), for several speaker positions, and using the $2P_{short}$ microphone arrangement. See Fig. 2 for geometrical references. (a) For position at angle 0°. (b) For position at angle 18°. (c) For position at angle 36°. (d) For position at angle 54°. (e) For position at angle 72°. (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

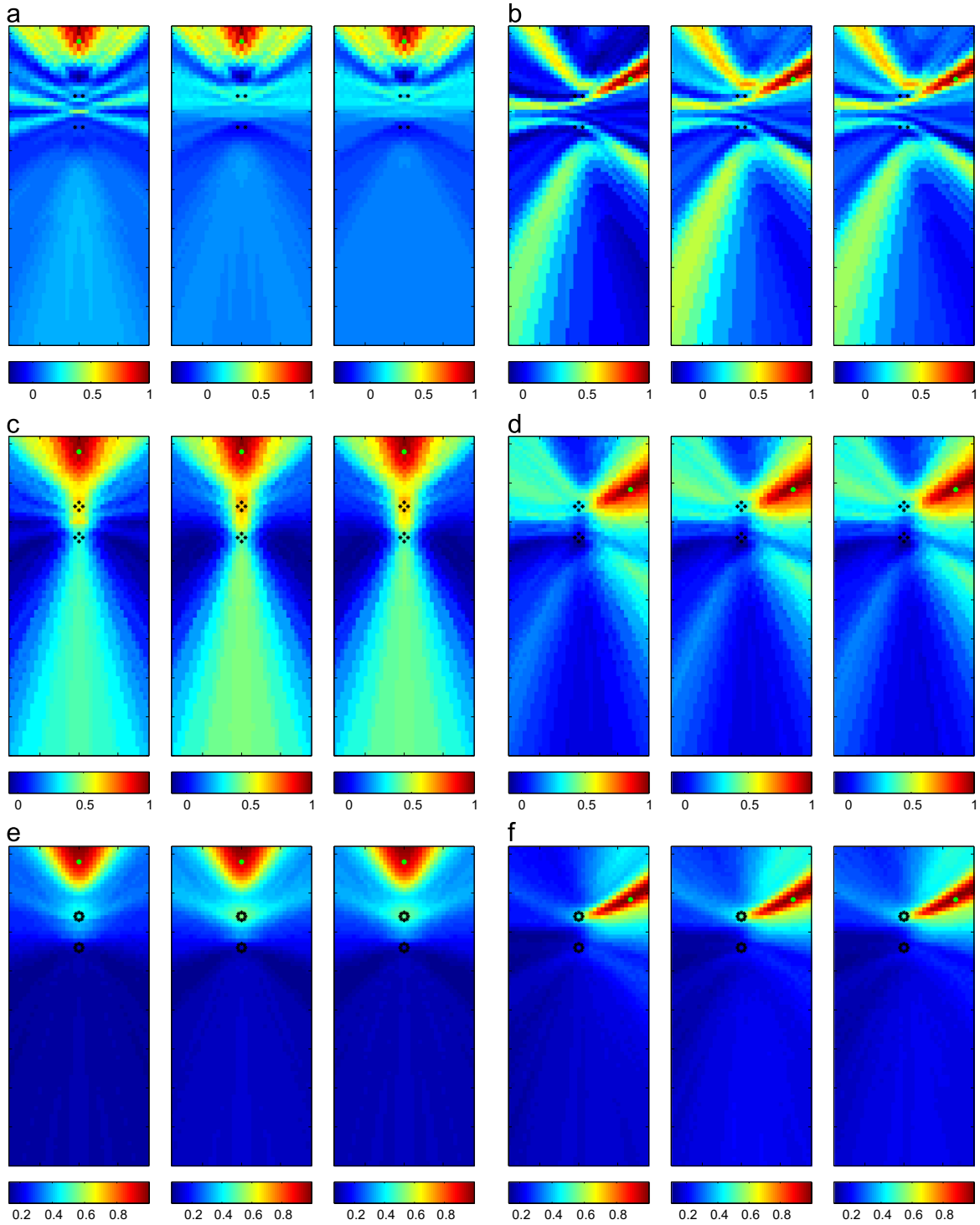


Fig. 6. Comparison between the SRP-PHAT map predicted by the model (left graphics), the simulated SRP-PHAT map (middle graphics), and the average (simulated) SRP-PHAT map (right graphics), for two speaker positions (at 0° and 72°), and using different microphone arrangements. See Fig. 2 for geometrical references. (a) For position at angle 0° ($2P_{long}$ arrangement). (b) For position at angle 72° ($2P_{long}$ arrangement). (c) For position at angle 0° ($6P$ arrangement). (d) For position at angle 72° ($6P$ arrangement). (e) For position at angle 0° (all pairs). (f) For position at angle 72° (all pairs). (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

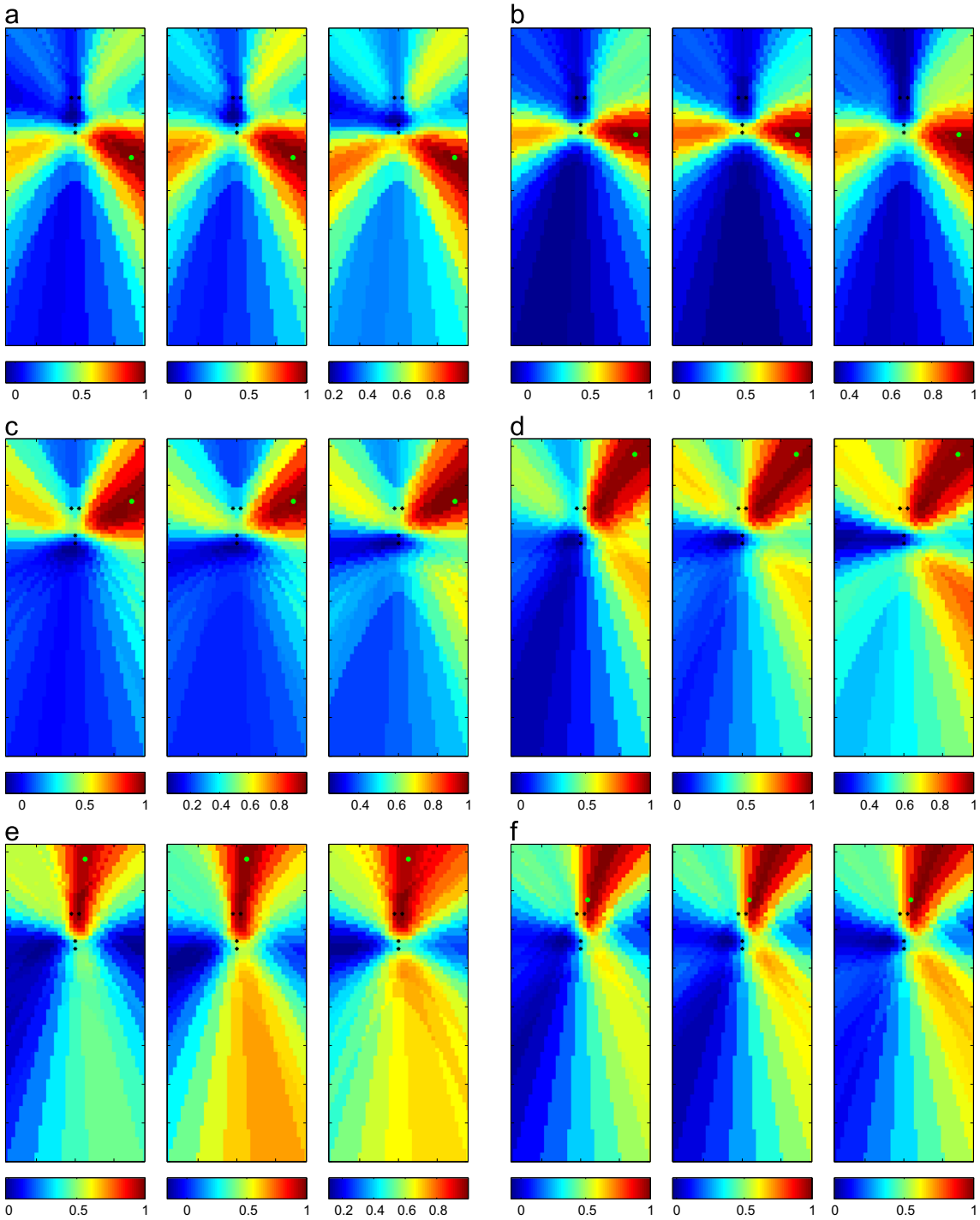


Fig. 7. Comparison between the SRP-PHAT map predicted by the model (left graphics), the real SRP-PHAT map (middle graphics), and the average (real) SRP-PHAT map (right graphics), for several speaker positions, and using the $2P_{short}$ microphone arrangement. See Fig. 2 for geometrical references. (a) For position 1. (b) For position 2. (c) For position 4. (d) For position 6. (e) For position 8. (f) For position 16. (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

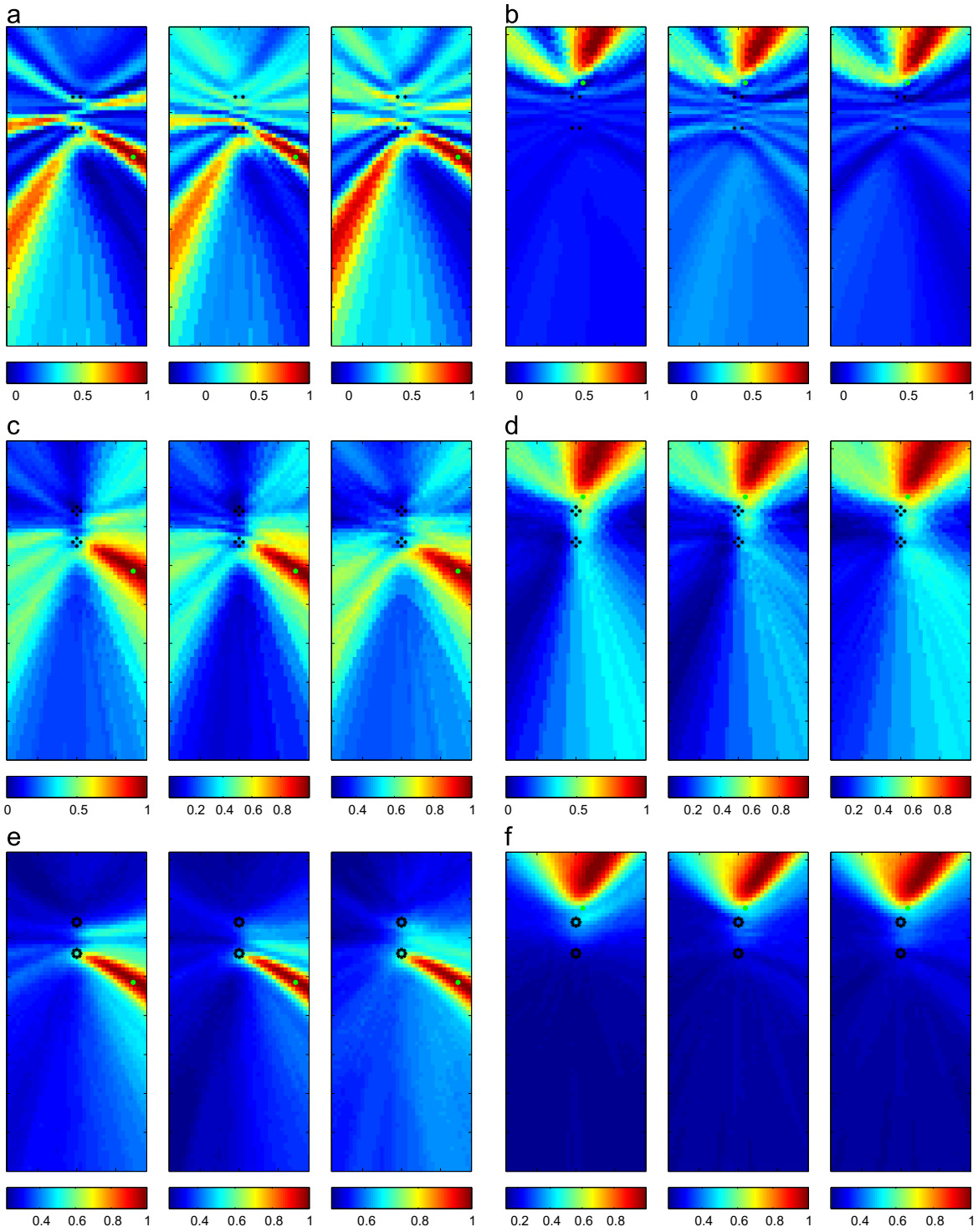


Fig. 8. Comparison between the SRP-PHAT map predicted by the model (left graphics), the real SRP-PHAT map (middle graphics), and the average (real) SRP-PHAT map (right graphics), for speaker positions 1 and 16, and using different microphone arrangements. See Fig. 2 for geometrical references. (a) For position 1 ($2P_{long}$ arrangement). (b) For position 16 ($2P_{long}$ arrangement). (c) For position 1 ($6P$ arrangement). (d) For position 16 ($6P$ arrangement). (e) For position 1 (all pairs). (f) For position 16 (all pairs). (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

frames in which the user was in the given position (for example, the right graphic in Fig. 5a).

In Fig. 5 we evaluated the results for the positions in the angles used in Figs. 3 and 4 (those emphasized in Fig. 2a), using two pairs of microphones (the $2P_{short}$ arrangement shown in Fig. 2c, composed of pairs $\mathbf{m}_1$–$\mathbf{m}_5$ and $\mathbf{m}_{11}$–$\mathbf{m}_{15}$), being orthogonal to each other, and belonging to the upper and lower arrays, respectively. This particular arrangement allows us to easily identify the expected hyperbolic shapes generated by each pair in the SRP-PHAT map (given a microphone pair, the place of points with equal time-difference of arrival is a hyperbola in the two-dimensional case, being a hyperboloid of revolution in the three-dimensional case), along with the sinc-like behaviour expressed by the proposed model. The figure shows an excellent agreement between the model predictions and the simulated SRP-PHAT maps.

In Fig. 6 we show the comparison considering different arrangements for the selected pairs of microphones (Fig. 6a and b for the $2P_{long}$ arrangement shown in Fig. 2c (using pairs $\mathbf{m}_1$–$\mathbf{m}_{13}$ and $\mathbf{m}_5$–$\mathbf{m}_9$), Fig. 6c and d for the 6*P* arrangement shown in Fig. 2c (using pairs $\mathbf{m}_1$–$\mathbf{m}_5, \mathbf{m}_3$–$\mathbf{m}_7, \mathbf{m}_9$–$\mathbf{m}_{13}, \mathbf{m}_{11}$–$\mathbf{m}_{15}, \mathbf{m}_1$–$\mathbf{m}_{13}$, and $\mathbf{m}_5$–$\mathbf{m}_9$), and Fig. 6e and f using *all* available pairs or microphones). In this case, only the positions at angles 0° and 72° are shown due to space constraints, and our target is showing how the model is able to accurately cope with very different microphone array topologies. Again, it can be clearly seen that the predictions are very similar to the simulated SRP-PHAT maps.

5.3. Validation with real data: SRP-PHAT prediction performance

In order to assess how accurate our model is in real scenarios, we have performed additional experiments using real recordings.

The experiments consisted on generating the SRP-PHAT acoustic power maps as predicted by the proposed model, and comparing it with the real SRP-PHAT acoustic power map calculated using the real acoustic waveforms from sequence 01 in the AV16.3 database. The IDIAP meeting room dimensions, with reflections up to order 2, have been considered for the model generation process, with the same acoustic assumptions discussed in the previous section. The evaluation includes the results of the comparison for several speaker positions (1, 2, 4, 6, 8 and 16, emphasized in Fig. 2b), that were selected to provide different acoustic situations, both in terms of distance and angular position with respect to the arrays.

In Fig. 7 we initially evaluated the results with the $2P_{short}$ arrangement for the microphone pairs (as in the evaluation shown in Fig. 5), to allow for an easy identification of the relevant effects (we again expect to see the hyperbolic components with the sinc-like behaviour predicted by the model). From Fig. 7, it can clearly be seen that, again, the predictions closely match the results with real data for the different acoustic conditions, even when the model predictions are generated using fixed and frequency independent average reflection coefficients, and that the acoustic model is based on the simplistic image method model.

Following the same approach than in the previous section, in Fig. 8 we next show the comparison considering different arrangements for the selected pairs of microphones (the same ones used for the evaluation shown in Fig. 6: $2P_{long}$, 6*P* and *all*), and only for positions 1 and 16 (due to space constraints). In this case, our target is again showing how the model deals with very different microphone array topologies. It can be clearly seen that even with the varied situations, the maps generated by the model accurately predict the real SRP-PHAT maps. It is specially interesting to note the results of the real maps for position 16 in Fig. 8, in which the maximum power value does not coincide at all with the speaker position (the green point). The fact that this is also correctly predicted by the model could be effectively used to improve speaker localization systems.

6. Model discussion

Given the parametric formulation of the proposed model, some relevant conclusions can be deduced from it. In this section, we will provide some discussion on the effects of signal bandwidth, distance between microphones, window size, and sampling frequency. For the sake of simplicity, we will only consider the pattern generated by a microphone pair, since the pattern of an array of any number of elements can be expressed as the sum of the contributions of all microphone pairs, as shown in Eq. (7). All the considerations discussed here can be easily extrapolated to any array topology.

6.1. Bandwidth considerations

The directivity patterns typically referred in the literature only take narrowband signals into account. This is not appropriate for signals such as speech where the *broadband* assumption is mandatory.

The proposed model, as well as the SRP formulation, integrates the response for the full signal bandwidth. This, along with the PHAT filtering, increases the immunity of the localization systems to the unknown spectral content of the acoustic signals.

Given the inverse relationship between frequency and wavelength, a better spatial resolution is expected for broader bandwidths (provided that the power in the higher frequencies is high enough compared with the noise). The proposed model is also consistent with this idea, and its analytic formulation contributes with two additional advantages. First, it allows us to derive constraints on the optimal bandwidth to use, provided the rest of system parameters, and, second, this optimization can be directly related to the final system performance, as the quality of the SRP-PHAT maps are directly related to their performance, so that optimization metrics could be derived from the model.

Fig. 9 shows the response power patterns predicted by the model for different bandwidths ($f_0 \in \{3\text{ kHz}, 4.5\text{ kHz}, 6\text{ kHz}\}$), where the increment in spatial resolution as the signal bandwidth increases can be easily seen.

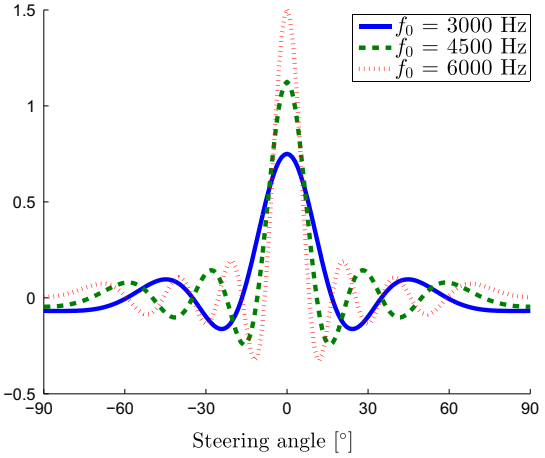


Fig. 9. Response Power Pattern for a microphone pair and different signal bandwidths ($D_{\mathbf{p}} = 20$ cm).

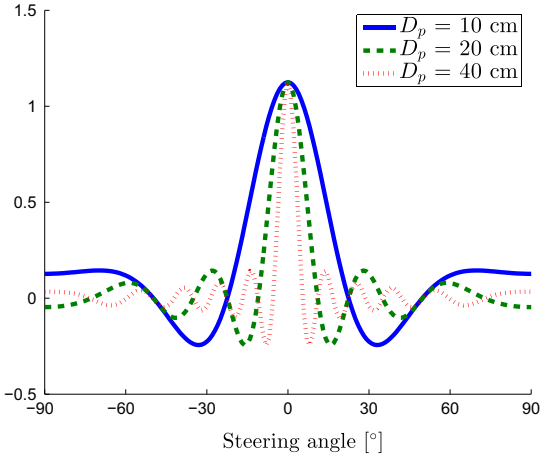


Fig. 10. Response Power Pattern for a microphone pair and different distances between microphones ($f_0 = 4.5$ kHz).

In Fig. 11, and for two speaker positions (1 and 6 this time), we show the model behavior as compared with the real data, for different bandwidths ($f_0 \in \{3\text{ kHz}, 4.5\text{ kHz}, 6\text{ kHz}\}$), and using the same type of graphics shown previously in Section 5.3. Again, there is a high agreement between the predicted maps and the real ones, and it is also clear the increment in spatial resolution for increasing bandwidths (the *width* of the hyperbolic regions gets narrower as the bandwidth increases), both for the model predictions and those calculated from the real data.

6.2. Distance between microphones

The distance between microphone elements is an important factor for the design of microphone arrays. Referring to Fig. 12, the time difference of arrival $\Delta\tau(\mathbf{p},\mathbf{x})$ between an acoustic source located at $\mathbf{x}$ and the microphones of a given pair $\mathbf{p}$ can be expressed as

$$\Delta\tau(\mathbf{p},\mathbf{r}) = \frac{1}{c}\left(\sqrt{R_{\mathbf{x}}^2 + d_{\mathbf{p}}^2 + D_{\mathbf{p}}R_{\mathbf{x}}\sin(\theta_{\mathbf{x}})} - \sqrt{R_{\mathbf{x}}^2 + d_{\mathbf{p}}^2 - D_{\mathbf{p}}R_{\mathbf{x}}\sin(\theta_{\mathbf{x}})}\right) \approx \frac{D_{\mathbf{p}}R_{\mathbf{x}}\sin(\theta_{\mathbf{x}})}{c\sqrt{R_{\mathbf{x}}^2 + d_{\mathbf{p}}^2}}, \tag{24}$$

where c is the sound velocity, $D_{\mathbf{p}} = 2d_{\mathbf{p}}$ is the distance between the microphones, $\theta_{\mathbf{x}}$ determines the angular position of $\mathbf{x}$ from the microphone pair center, and $R_{\mathbf{x}}$ is the distance from this center to $\mathbf{x}$. The final approximation uses the Taylor Series $\sqrt{1+x} = 1 + \frac{x}{2} + \cdots$ to simplify the equation.

In this section we will analyze the influence of the distance between microphones, as compared to the distance to the speaker, in the received power pattern. Two cases attending the source position are discussed, to finally provide some insights on the spatial aliasing effects.

6.2.1. Far-field considerations

When the distance between the speaker and the array is much higher than the distance between both microphones[2] ($R_{\mathbf{x}} \gg d_{\mathbf{p}}$), the power pattern does not depend on how far the speaker is, as from Eq. (24):

$$\Delta\tau(\mathbf{p},\mathbf{r}) \approx \frac{D_{\mathbf{p}}\sin(\theta_{\mathbf{x}})}{c}, \tag{25}$$

so that the power pattern will only depend on the arrival angles, and the angular resolution could be improved by increasing the separation between the microphones (see Section 6.2.2 below, for details on spatial aliasing).

Using Eq. (25) in Eq. (22), we can get the expression of the power pattern in far-field conditions:

$$P_{FF}(\mathbf{q}) = P_{FF}(\theta_{\mathbf{q}}) = \frac{4\sin\left(\omega_0\frac{D_{\mathbf{p}}}{c}(\sin(\theta_{\mathbf{q}}) - \sin(\theta_{\mathbf{r}}))\right)}{\frac{D_{\mathbf{p}}}{c}(\sin(\theta_{\mathbf{q}}) - \sin(\theta_{\mathbf{r}}))}, \tag{26}$$

where $\theta_{\mathbf{q}}$ and $\theta_{\mathbf{r}}$ are the array steering and arrival angles respectively.

Fig. 10 shows the predicted response power patterns for different distances between microphones ($D_{\mathbf{p}} \in \{10\text{ cm}, 20\text{ cm}, 40\text{ cm}\}$), in the far-field case, in which we can see how the angular resolution improves for increasing $D_{\mathbf{p}}$.

On the other hand, when $R_{\mathbf{x}} \ll d_{\mathbf{p}}$, the power pattern can discriminate angle as well as distance, and the distance between microphones is not relevant anymore as, from Eq. (24):

$$\Delta\tau(\mathbf{p},\mathbf{r}) \approx \frac{2R_{\mathbf{x}}\sin(\theta_{\mathbf{x}})}{c}. \tag{27}$$

Using Eq. (27) in Eq. (22), we can get the expression of the power pattern in this condition:

$$P_{NF}(\mathbf{q}) = P_{NF}(R_{\mathbf{q}},\theta_{\mathbf{q}}) = \frac{2c\sin\left(2k_0(R_{\mathbf{q}}\sin(\theta_{\mathbf{q}}) - R_{\mathbf{r}}\sin(\theta_{\mathbf{r}}))\right)}{R_{\mathbf{q}}\sin(\theta_{\mathbf{q}}) - R_{\mathbf{r}}\sin(\theta_{\mathbf{r}})}, \tag{28}$$

where $k_0 = \frac{\omega_0}{c}$ is the wavenumber at ω_0. Eq. (28) shows that the power pattern depends on the projection of the

[2] The theoretical far-field criterion is $R_{\mathbf{x}} > \frac{\pi D_{\mathbf{p}}^2}{\lambda}$ [28], but we have made the ($R_{\mathbf{x}} \gg d_{\mathbf{p}}$) assumption for simplicity, as this is not a critical limit for our considerations.

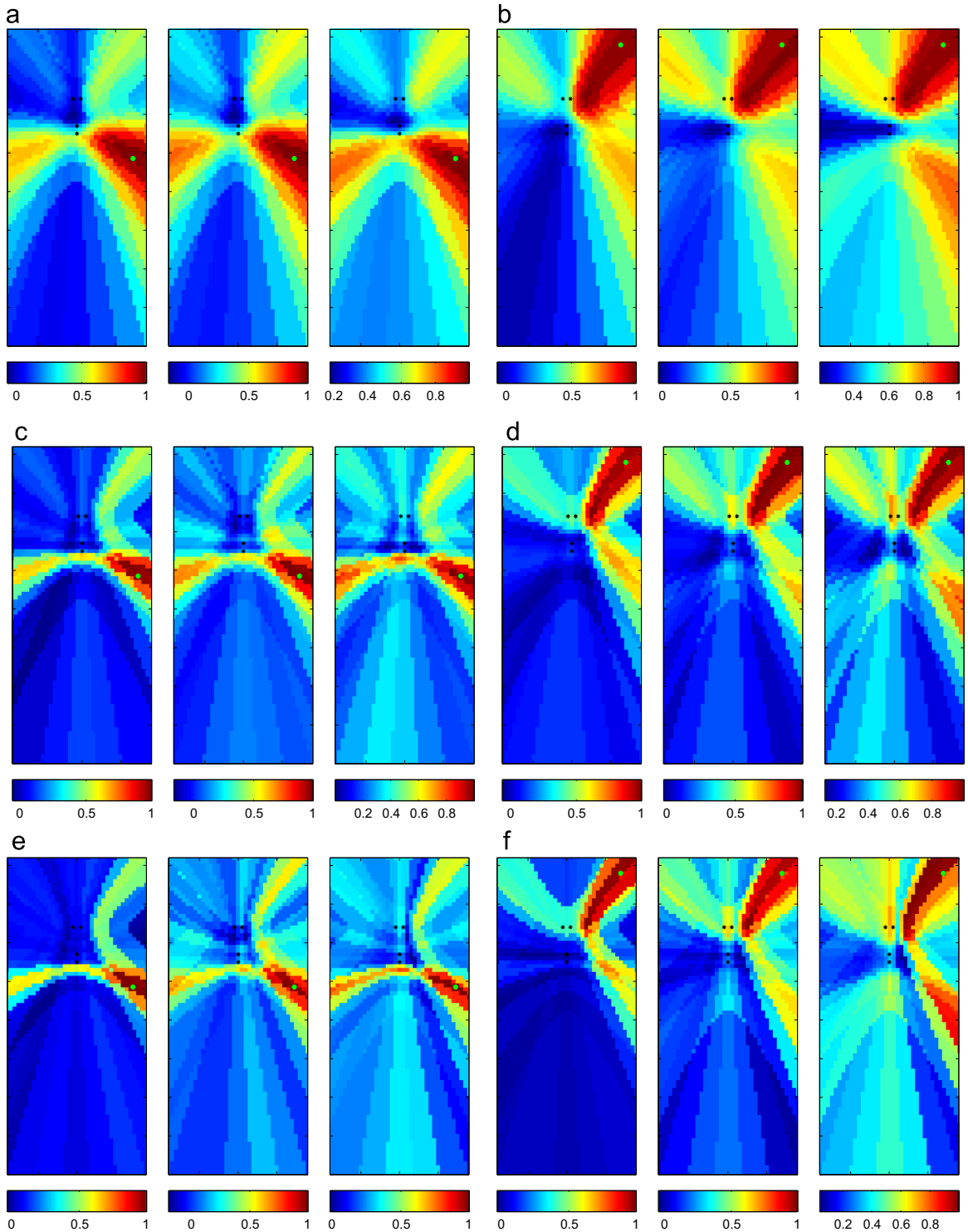


Fig. 11. Comparison between the SRP-PHAT map predicted by the model (left graphics), the real SRP-PHAT map (middle graphics), and the average (real) SRP-PHAT map (right graphics), for different bandwidths and two speaker positions. See Fig. 2 for geometrical references. (a) For position 1 ($f_0 = 1.5$ kHz). (b) For position 6 ($f_0 = 1.5$ kHz). (c) For position 1 ($f_0 = 3$ kHz). (d) For position 6 ($f_0 = 3$ kHz). (e) For position 1 ($f_0 = 4.5$ kHz). (f) For position 6 ($f_0 = 4.5$ kHz). (For a better visualization of the color information in these figures, the reader is referred to the web version of this paper.)

source and steered positions on the line linking the corresponding microphones ($R_{\mathbf{r}}\sin(\theta_{\mathbf{r}})$ and $R_{\mathbf{q}}\sin(\theta_{\mathbf{q}})$).

In Fig. 13, and for two speaker positions (again 1 and 6), we show the model behavior using two microphone pairs, as compared with the real data (with the same graphical information than the one shown in Section 5.3), and for two microphone distances: 20 cm (corresponding to microphone pairs $\mathbf{m}_1$–$\mathbf{m}_5$ and $\mathbf{m}_{11}$–$\mathbf{m}_{15}$ in AV16.3, as shown in Fig. 2c), and 82.46 cm (which was selected as the most suitable given the array topology, and corresponding to pairs $\mathbf{m}_1$–$\mathbf{m}_{13}$ and $\mathbf{m}_5$–$\mathbf{m}_9$, as shown in Fig. 2c). Again, there is a high agreement between the predicted maps and the real ones, and it is also clear the increment in spatial resolution for increasing distance between microphones (the *width* of the hyperbolic regions gets narrower as $D_{\mathbf{p}}$ increases).

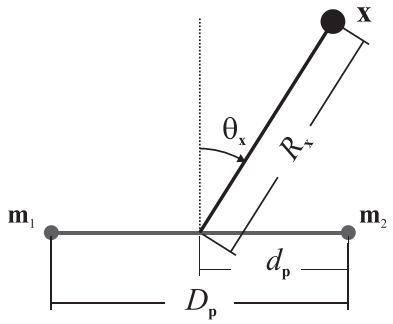


Fig. 12. Geometrical details for Section 6.2.

6.2.2. Spatial aliasing

The microphone array theory states that spatial aliasing occurs when $D_{\mathbf{p}} > \frac{\lambda}{2}$. When it happens, grating lobes appear in the array directivity pattern, negatively affecting the array spatial discrimination.

Some authors have pointed out that the spatial Nyquist criterion has limited importance for microphone arrays [37], and specifically when considering broadband steered response patterns [38]. The model proposed in this paper is also able to explain this characteristic.

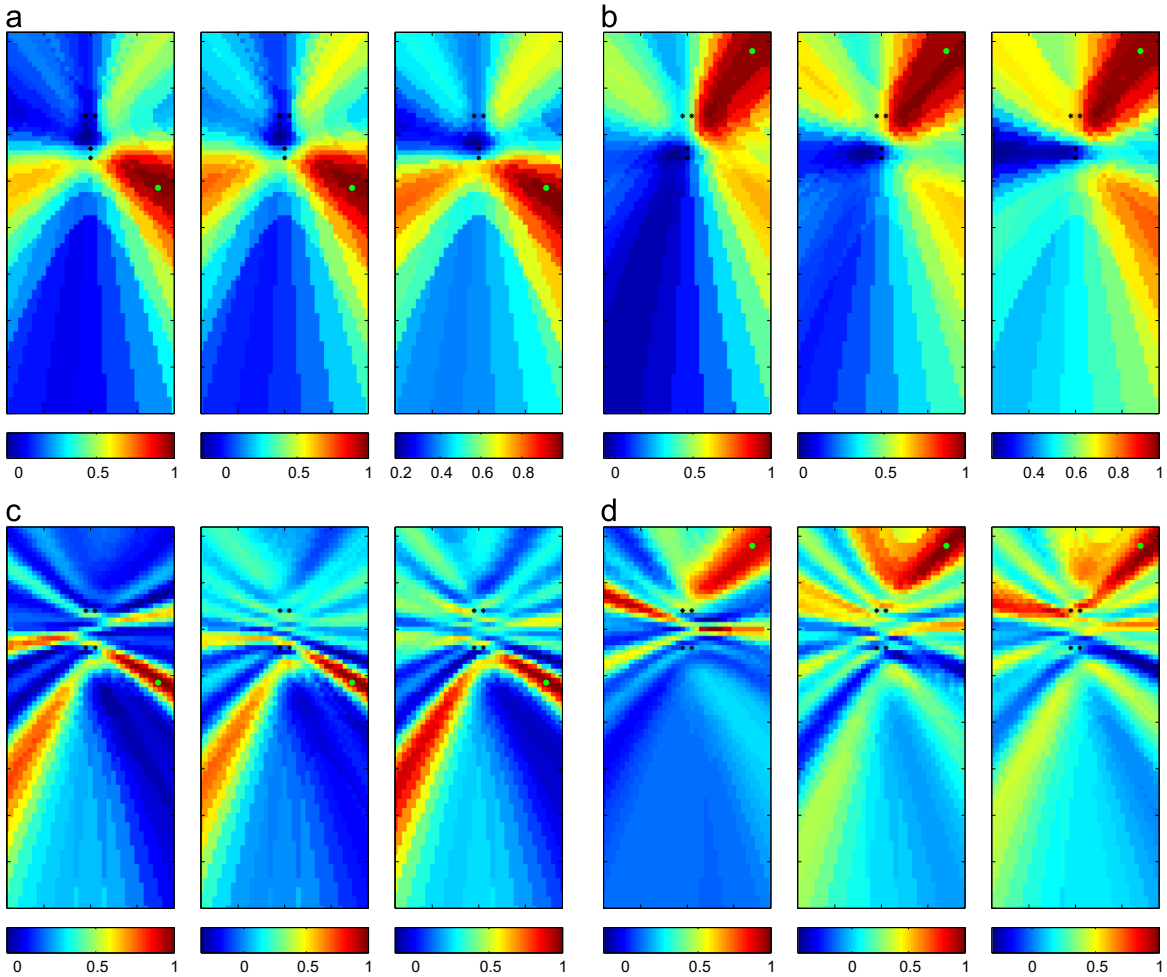


Fig. 13. Comparison between the SRP-PHAT map predicted by the model (left graphics), the real SRP-PHAT map (middle graphics), and the average (real) SRP-PHAT map (right graphics), for different microphone spacings and two speaker positions ($f_0 = 1.5$ kHz). See Fig. 2 for geometrical references. (a) For position 1 ($D_{\mathbf{P}} = 20$ cm). (b) For position 6 ($D_{\mathbf{P}} = 20$ cm). (c) For position 1 ($D_{\mathbf{P}} = 82.46$ cm). (d) For position 6 ($D_{\mathbf{P}} = 82.46$ cm).

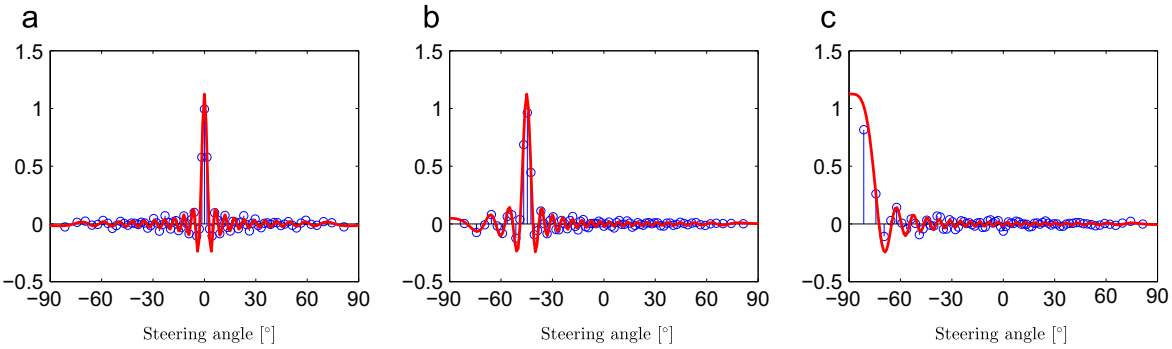


Fig. 14. Microphone pair power pattern in an anechoic simulation, showing no spatial aliasing effects. $D_{\mathbf{p}} = 82.46$ cm, aliased frequencies start at 206 Hz and signal bandwidth is $f_0 = 1.5$ kHz. (a) 0°. (b) −45°. (c) −90°.

For narrowband signals, the main lobe of the power pattern always appears in the steered direction, but the position of the grating lobes changes with frequency. For a broadband signal, the use of the SRP-PHAT approach implies an integration over the full bandwidth, and equally weighting all frequency components (due to the use of the PHAT filtering). That way, the main lobes are coherently merged, while the grating lobes are incoherently combined, thus reducing the impact of spatial aliasing.

Fig. 14 shows a microphone pair power pattern for three different directions (0°, −45° and −90°) when the microphones are separated $D_{\mathbf{p}} = 82.46$ cm (same separation as the one shown for pairs in Fig. 2c). For this $D_{\mathbf{p}}$, aliasing should occur for frequencies above 206 Hz, well below the considered bandwidth ($f_0 = 1.5$ kHz). The model prediction and the simulated waveforms were generated in the same conditions than Section 5.1, except for the assumption of anechoic propagation (in order to avoid other effects). Neither the simulated SRP-PHAT power pattern (shown as *stems*) nor the model prediction (shown as a *solid line*, which perfectly fits the simulated SRP-PHAT) exhibits aliasing effects, and only the main lobe is present in both cases.

6.3. Signal and FFT window lengths

The temporal signal window length, $T_W < \frac{N_F}{2f_s}$, is also an important parameter in real implementations of the SRP-PHAT algorithm. It determines the length of the analyzed audio segment, the computational demands, and the response time of the system. The window length is, as shown in Section 3.6, also related to the FFT size (N_F).

Geometrically, the minimal required window length is determined by the maximum distance between microphones ($T_W > \frac{2D_{\mathbf{p}}}{c}$), in order to allow calculating all the possible steering delays. However, in practical discrete time implementations using signal windows, this condition should become $T_W \gg \frac{D_{\mathbf{p}}}{c}$, because if this is not satisfied for some microphone pair, the signal segment captured by a microphone cannot be approximated as a simply time-shifted instance of the segment captured by the other microphone, as was assumed in Eqs. (8) and (13). This limitation is usually not considered in the literature, as for typical microphone arrays, the window length is long enough. However, for applications combining microphones with high $D_{\mathbf{p}}$ in spatially distributed microphone arrays, it should be considered.

With the previous restrictions in the values of T_W, it follows that $N_F \gg \frac{2D_{\mathbf{p}} f_s}{c}$. This expression also guarantees the assumption made in Section 3.6 related to the fact that the discrete and continuous versions of the model (Eqs. (22) and (9)) are proportional when N_F is high enough (as the restriction in N_F allows to apply the small angle approximation ($\sin(x) \approx x$,for $x \to 0$) in the denominator of Eq. (22)).

Finally, longer window lengths may also lead to a worse spectral estimation for moving sources, as the results will integrate effects for different source positions.

7. Conclusions and future work

In this paper we have proposed an analytical model that accurately predicts the acoustic power maps generated by the SRP-PHAT algorithm in both anechoic and non-anechoic conditions. It is based on reasonable assumptions about sound propagation and its interaction with the environment. Our model predicts that SRP-PHAT (and the corresponding GCC-PHAT functions from which SRP-PHAT is calculated) depends on the topology of the array, room's geometry and signal bandwidth, but not on the spectral content of the signal. This last property is very important in speaker localization as the speech signal is unknown. Our model allows us to discuss the influence of all these factors in the localization accuracy, specially in reverberant scenarios. The model has been thoroughly validated using simulated and real data for a wide range of conditions (speaker positions, bandwidth and array topology considerations). In the synthetic case we show that our model predictions are very close to that provided by the image method, a standard room acoustics simulator. We also tested our model with real data from a publicly available dataset. Our results are reproducible and verify empirically that the model is able to reproduce SRP-PHAT power maps with high fidelity in a real case.

Regarding potential applications, the proposed model is parametric, analytical and differentiable in function of all aforementioned factors and the position of an acoustic source. We thus believe that this model can be of high interest to improve SRP-PHAT (and GCC-PHAT) based speaker localization in several ways. Given that the

geometry of the room is known, our model can predict the effect of reverberation and can serve as a way to avoid sampling the possible space of target positions. Also, as the array topology is an input parameter of our model, it can play an important role for improving methods that find optimal array topologies [39,40] (for instance maximizing the predicted accuracy), and in automatic array geometry calibration systems [41,25,42]. In a similar case, given that the room's geometry is also an input parameter, we believe that our model can be used to contribute to the automatic identification of room's geometry [43–45] from acoustic measurements.

Regarding future work, we first plan to apply the proposal to the localization system described in [46], that will be clearly benefited, as it is based on the use of a generative model with sparse constraints. Additionally, and taking into account that the actual success of the model application depends on the availability of the room geometrical details and the array topology, the integration of the proposal with automatic calibration and room geometry estimation strategies will also be addressed.

Acknowledgments

This work has been supported by the Spanish Ministry of Economy and Competitiveness under Project SPACES-UAH (TIN2013-47630-C2-1-R), and by the FPU Grants Program of the University of Alcalá.

References

[1] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282.
[2] M. Marron-Romera, J. Garcia, M. Sotelo, D. Pizarro, M. Mazo, J. Cañas, C. Losada, A. Marcos, Stereo vision tracking of multiple objects in complex indoor environments, Sensors 10 (10) (2010) 8865–8887.
[3] M.S. Brandstein, H.F. Silverman, A practical methodology for speech source localization with microphone arrays, Comput. Speech Lang. 11 (2) (1997) 91–126, http://dx.doi.org/10.1006/csla.1996.0024.
[4] J. DiBiase, H. Silverman, M. Brandstein, Robust localization in reverberant rooms, Microphone Arrays, 2001, pp. 157–180.
[5] A. Waibel, R. Stiefelhagen, Computers in the Human Interaction Loop, 1st ed., Springer Publishing Company, Incorporated, Berlin Heidelberg, 2009.
[6] J. DiBiase, A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays (Ph.D. thesis), Brown University, 2000.
[7] M. Gillette, H. Silverman, A linear closed-form algorithm for source localization from time-differences of arrival, IEEE Signal Process. Lett. 15 (2008) 1–4, http://dx.doi.org/10.1109/LSP.2007.910324.
[8] B. Mungamuru, P. Aarabi, Enhanced sound localization, IEEE Trans. Syst. Man Cybern. Part B 34 (3) (2004) 1526–1540.
[9] S.T. Birchfield, A unifying framework for acoustic localization, In: Proceedings of the 12th European Signal Processing Conference (EUSIPCO), 2004.
[10] H.T.H. Do, Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array (Ph.D. thesis), Brown University, 2011.
[11] M. Brandstein, D. Ward (Eds.), Microphone Arrays : Signal Processing Techniques and Applications, Springer, London, 2001.
[12] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. 24 (4) (1976) 320–327, http://dx.doi.org/10.1109/TASSP.1976.1162830.
[13] C. Zhang, D. Florencio, Z. Zhang, Why does PHAT work well in low noise, reverberative environments?, in: Proceedings of ICASSP 2008, 2008, pp. 2565 –2568, http://dx.doi.org/10.1109/ICASSP.2008.4518172.
[14] M. Omologo, P. Svaizer, Use of the cross-power-spectrum phase in acoustic event location, IEEE Trans. Speech Audio Process. 5 (1993) 288–292.
[15] H. Buchner, R. Aichner, W. Kellermann, TRINICON-based blind system identification with application to multiplesource localization and separation, Blind Speech Separation, Springer-Verlag, Netherlands, 2007, 101–147.
[16] P. Pertilä, M. Hämäläinen, A track before detect approach for sequential Bayesian tracking of multiple speech sources, In: Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing, 2010, pp. 4974–4977.
[17] H.F. Silverman, Y. Yu, J.M. Sachar, W.R. Patterson, Performance of real-time source-location estimators for a large-aperture microphone array, IEEE Trans. Speech Audio Process. 4 (13) (2005) 593–606.
[18] J.P. Dmochowski, J. Benesty, Steered beamforming approaches for acoustic source localization, In: I. Cohen, J. Benesty, S. Gannot (Eds.), Speech Processing in Modern Communication, Springer Topics in Signal Processing, vol. 3, Springer, Berlin, Heidelberg, 2010, pp. 307–337, http://dx.doi.org/10.1007/978-3-642-11130-3_12.
[19] J. Dmochowski, J. Benesty, S. Affes, A generalized steered response power method for computationally viable source localization, IEEE Trans. Audio Speech Lang. Process. 15 (8) (2007) 2510–2526, http://dx.doi.org/10.1109/TASL.2007.906694.
[20] A. Badali, J.-M. Valin, F. Michaud, P. Aarabi, Evaluating real-time audio localization algorithms for artificial audition in robotics, In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, IROS 2009, 2009, pp. 2033 –2038, http://dx.doi.org/10.1109/IROS.2009.5354308.
[21] H. Do, H. Silverman, SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data, In: Proceedings of ICASSP 2010, 2010, pp. 125–128, http://dx.doi.org/10.1109/ICASSP.2010.5496133.
[22] M. Cobos, A. Marti, J. Lopez, A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling, IEEE Signal Process. Lett. 18 (1) (2011) 71–74.
[23] T. Butko, F. Gonzalez Pla, C. Segura Perales, C. Nadeu Camprubí, F.J. Hernando Pericás, Two-source acoustic event detection and localization: online implementation in a smart-room, In: Proceedings of the 17th European Signal Processing Conference (EUSIPCO'11), 2011, pp. 1317–1321.
[24] K.D. Donohue, J. Hannemann, H.G. Dietz, Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments, Signal Process. 87 (7) (2007) 1677–1691.
[25] S. Khanal, H.F. Silverman, R.R. Shakya, A free-source method (FrSM) for calibrating a large-aperture microphone array, IEEE Trans. Audio Speech Lang. Process. 21 (8) (2013) 1632–1639.
[26] C. Zhang, D. Florencio, D. Ba, Z. Zhang, Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings, IEEE Trans. Multimedia 10 (3) (2008) 538–548, http://dx.doi.org/10.1109/TMM.2008.917406.
[27] G.C. Carter, A.H. Nuttall, P. Cable, The smoothed coherence transform, Proc. IEEE 61 (10) (1973) 1497–1498.
[28] L.J. Ziomek, Fundamentals of Acoustic Field Theory and Space–Time Signal Processing, CRC Press, Boca Raton, Florida, 1995.
[29] S.T. Neely, J.B. Allen, Invertibility of a room impulse response, J. Acoust. Soc. Am. 66 (1) (1979) 165–169.
[30] B.F.G. Katz, F. Prezat1, C. d'Alessandro, Human voice phoneme directivity pattern measurements, J. Acoust. Soc. Am. 120 (5) (2006).
[31] G. Lathoud, J.-M. Odobez, D. Gatica-Perez, AV16.3: an audio–visual corpus for speaker localization and tracking, In: S. Bengio, H. Bourlard (Eds.), Proceedings of the MLMI, Lecture Notes in Computer Science, vol. 3361, Springer-Verlag, Berlin Heidelberg, 2004, pp. 182–195.
[32] D.C. Moore, The IDIAP Smart Meeting Room, Technical Report, IDIAP Research Institute, Switzerland, 2004.
[33] G. Lathoud, AV16.3 Dataset ⟨http://www.idiap.ch/dataset/av16-3/⟩, last accessed in October 2013.
[34] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am. 65 (4) (1979) 943–950, http://dx.doi.org/10.1121/1.382599.
[35] ULYSSES Absorber Database Download ⟨http://www.ifbcon.de/software/ulysses/downloads/abs/e.php⟩, last accessed in October 2013.
[36] W.C. Sabine, Collected Papers on Acoustics, Cambridge Harvard University Press, Dover, New York, 1922 (1964 reprint).
[37] G. Lathoud, Spatio-temporal analysis of spontaneous speech with microphone arrays (Ph.D. thesis), École Polytechnique Fédérale de

Lausanne, Lausanne, Switzerland, Ph.D. Thesis #3689 at the École Polytechnique Fédérale de Lausanne (IDIAP-RR 06-77), 2006.
[38] J. Dmochowski, J. Benesty, S. Affès, On spatial aliasing in microphone arrays, Trans. Signal Process. 57 (4) (2009) 1383–1395, 10.1109/TSP.2008.2010596.
[39] Z. Li, R. Duraiswami, Flexible and optimal design of spherical microphone arrays for beamforming, IEEE Trans. Audio Speech Lang. Process. 15 (2) (2007) 702–714.
[40] I. Kodrasi, T. Rohdenburg, S. Doclo, Microphone position optimization for planar superdirective beamforming, in: Proceedings of ICASSP 2011, IEEE, Prague, Czech Republic, 2011, pp. 109–112.
[41] M.J. Taghizadeh, P.N. Garner, H. Bourlard, Enhanced diffuse field model for ad hoc microphone array calibration, Signal Process. 101 (2014) 242–255. ISSN 0165-1684.
[42] N.D. Gaubitch, W.B. Kleijn, R. Heusdens, Auto-localization in ad-hoc microphone arrays, in: Proceedings of ICASSP 2013, 2013, pp. 106–110.
[43] A.H. Moore, M. Brookes, P.A. Naylor, Room geometry estimation from a single channel acoustic impulse response, in: Proceedings of European Signal Processing Conference (EUSIPCO), Marrakech, Morocco, 2013.
[44] A. Asaei, M. Golbabaee, H. Bourlard, V. Cevher, Room acoustic modeling exploiting joint sparsity and low-rank structures, In: Signal Processing with Adaptive Sparse Structured Representations SPARS, 2013.
[45] S. Tervo, T. Tossavainen, 3D room geometry estimation from measured impulse responses, In: Proceedings of ICASSP 2012, IEEE, Kyoto, Japan, 2012, pp. 513–516.
[46] J. Velasco, D. Pizarro, J. Macias-Guarasa, Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints, Sensors 12 (2012) 13781–13812.

Chapter 5

Technical Report on Sparse Acoustic Source Localization: *Acoustic Localization Based on an GCC-PHAT Generative Model and Sparse Optimization Techniques*

Acoustic Localization Based on an GCC-PHAT Generative Model and Sparse Optimization Techniques

Jose Velasco[1], Daniel Pizarro[1], and Javier Macias-Guarasa[1]

[1]Department of Electronics, Escuela Politécnica Superior, University of Alcalá, 28805 Alcalá de Henares, Spain

November 22, 2016

Abstract

This paper presents a novel approach for indoor acoustic multi-source localization using sensor arrays. This new approach is based on the reliable assumption that only few punctual sources are active at the same moment, and the coherence between them is small. The position of the active sources is determined by fitting the measured GCC-PHAT to a model imposing sparse constraints. The employed model has been previously described in the literature for a single source, this paper proves that it behaves linearly for multiple sources if small coherence between them is assumed

1 Introduction

The acoustic source localization methods are the starting point of other techniques like voice enhancing using beamforming. Therefore, acoustic source localization has received significant attention lately as a mode of automatic tracking of persons, and as a complement to other existing alternatives of tracking, e.g. the CHIL (Computer in Human Interaction Loop) project [45]

Many approaches exist in the literature, all of them using microphone arrays as a non intrusive audio acquisition method. These can roughly be divided in three categories [6,18]: *i)* two-stage time delay estimation based methods, *ii)* one-stage beamforming based methods, and *iii)* high-resolution spectral-estimation based methods. Methods in *iii)* are not able to efficiently cope with real-world conditions (mainly noise and reverberation issues), making *i)* and *ii)* the leading methods.

Methods in *i)* are composed of two stages: in the first one they estimate the time-difference of arrival (TDOA) of signals between pairs of microphones [5]. This is usually done using generalized cross-correlation (GCC) techniques [23]. Among the possible weighting functions, the Phase Transform (PHAT) has been found to perform very well under realistic acoustical environments [46], leading to the GCC-PHAT [23] (also known as the Crosspower-Spectrum Phase [29]). There are also alternative methods, such as those based on Blind Source Separation [7], or those using a likelihood function of phase differences [30]. In a second step, the TDOA information is combined with knowledge of the microphones' positions, using optimization techniques (maximum likelihood, least squares, spherical interpolation, etc.), to generate a spatial estimator of the source position [6,18,33]. The methods in *i)* are usually not well suited to multisource scenarios, and their main problem is their sensitivity to errors in the TDOA estimation, that can be hardly corrected if severe enough [17,44].

On the other hand, beamforming [19] based techniques estimate the position of the source by sampling a set of possible spatial locations and computing a beamforming function at each location. The approach then chooses the source location that maximizes a statistic that is maximum when the target position matches the source location. For instance in SRP, which is the simplest beamforming method, the statistic is based on the signal power received when the microphone array is steered in the direction of a specific location. SRP-PHAT is a widely used algorithm for speaker localization based on beamforming. It was first proposed in [17] [1] and is a beamforming based method that combines the robustness of the steered beamforming methods with the insensitivity to signal conditions afforded by the PHAT filter. The classical *delay-and-sum* beamformer used in SRP is replaced in SRP-PHAT by a *filter-and-sum* beamformer using PHAT filtering to weight the incoming signals.

The main problem with beamforming methods is their high computational cost, provided that they have to sample all potential positions of the space, labeling all local maxima as position candidates for acoustic sources.

This paper focuses on audio-based localization in a very general scenario, where unknown wide-band audio sources (e.g. human voice) are captured by a set of microphone arrays placed in known positions. The main objective of the paper is to use the GCC-PHAT correlations computed from the signals captured by the microphone arrays to robustly obtain the position of the active acoustic sources.

We propose an optimization approach to fit the generative model proposed in [42] to noisy GCC-PHAT data. We exploit the fact that only a few speakers are expected to be active at the same time. This simple idea is modeled using sparse constraints in the optimization task.

The proposed approach, unlike methods in *i)*, takes advantage of all the valuable information in GCC-PHAT. It has the same philosophy as the beamforming based methods, which combines the GCC-PHAT measures in a robust manner, but since we use a generative model, our approach is able to perform hiper-resolution even when only few microphones are available. Furthermore, the proposed method is well suited for multi-source scenarios.

This paper has a limited experimental section where it is shown that our model-based approach has results similar to SRP-PHAT in a single source-scenario but improving the fine error. In multi-source scenarios, we also show some qualitative results of the promising behavior of this approach.

1.1 Notation

Real scalar values are represented by lowercase letters (e.g. δ). Vectors are by default arranged column-wise and are represented by lowercase bold letters (e.g. $\mathbf{x}$). Matrices are represented by uppercase bold letters (e.g. $\mathbf{M}$). Upper-case letters are reserved to define vector and set sizes (e.g. vector $\mathbf{x} = (x_1, \cdots, x_N)^\top$ is of size N), and $\mathbf{x}^\top$ denotes transpose of vector $\mathbf{x}$. The l_2 norm of a vector $\|\mathbf{x}\|_2 = \left(|x_1|^2 + \cdots + |x_N|^2\right)^{\frac{1}{2}}$ will be written by default as $\|.\|$ for simplicity. Calligraphic fonts are reserved to represent ranges or sets (*e.g.*, $\mathbb{R}$ for real or generic sets $\mathcal{G}$). Continuous time signals are represented by scalar functions of t variable as for instance $x(t)$. Discrete time signals use k to denote discrete time samples. The Fourier transform of a continuous signal $x(t)$ is represented with complex function $X(\omega)$, with $X^*(\omega)$ being the complex-conjugate of $X(\omega)$ and $|X(\omega)| = \sqrt{X^*(\omega)X(\omega)}$. $\mathrm{Re}(X(\omega))$ and $\mathrm{Im}(X(\omega))$ are the real and imaginary parts of $X(\omega)$ respectively. We refer to supp(.) as the support function.

[1] Although the formulation is virtually identical to the *Global Coherence Field* (GCF) described in [29].

1.2 Paper Structure

The rest of paper is distributed as follows. In section 2 we provide the state-of-the-art in sparse representation of signal an sparse source localization. Section 3 recall those acoustic source localization based on GCC.In section 4 we describe the redundancy of the previous methods when PHAT filtering is in use. We extend the model proposed in [42] to the multisource case. In section 6 we propose new algorithms for acoustic source localization using sparse constrains. We also provide some experimentation to validate the proposed algorithms using real data in section 7. Finally, conclusions are drawn in section 8.

2 State of the art

2.1 Sparse Representation of Signals

Many areas of science share the principle of parsimony as the central criterion: the simplest explanation of a given phenomenon is preferred over more complicated ones. This brilliant idea has been recently applied to the representation of signals using overcomplete basis sets, sometimes called dictionaries in the machine learning discipline. As a difference with respect to traditional basis functions (e.g. Fourier basis functions), overcomplete dictionaries have more degrees of freedom than those necessary to represent the signal. The mathematical tool to impose parsimony in the representation of a signal, when several choices are available, is given by imposing the so-called sparse constraints. The basic idea is to use the lowest amount of coefficients to represent a signal with the basis functions. Sparse constraints, if applicable, allow to beat up several theoretical barriers in signal compression and representation [4, 9].

The main way sparsity is imposed is by using optimization approaches, where the l_0 norm is the usual way to impose sparsity to vectors [9].

Most of the problems in which sparsity is included using the l_0 norm are very difficult to solve. Several methods have been proposed to find sparse representations, including brute force approaches as well as more computationally efficient approximate methods such as 'non linear programming' [31], and greedy pursuit [16, 36, 38]. Among all approximate solutions, l_1 norm based convex relaxations have flourished in the literature. Among them, we can be highlight the Basis Pursuit method [13,39], originally introduced by [14] almost 40 years ago, but more recently revisited with a profound theoretical study in the past decade, due to its intensive use in the modern compressive sensing techniques [4, 9]. These methods provide very effective polynomial time algorithms that, under certain circumstances, are even equivalent to the original l_0 based problems [9, 39].

2.2 Sparse Source Localization

In last few years, sparse techniques explained above have been applied to the source localization problem in very different fashions.

In [26, 27], a localization approach based on sensor arrays is proposed. The signal obtained in each sensor is expressed as a linear combination of an attenuated and phase shifted version of the original and known signals emitted by the source. These conditions form an overcomplete linear model, where, thanks to sparse constraints, the position of the sources is given. Also in [26, 27] they propose to use *singular value decomposition* (SVD) to reduce problem size and filter noise in problems using multiple time samples.

Numerous modifications of the ideas proposed in [26] have been further developed. For example, in [35] an adaptive algorithm to dynamically adjust both the overcomplete basis and

the sparse solution is proposed. Also, the concept of Compressive Sensing [9] has been used in order to carry out distributed localization, reducing the information transmitted between sensors. Nevertheless, the sparse source localization algorithms discussed above, don't perform well and are not properly tested in real acoustic reverberant environments due to the input signals coherence caused by multipath.

In acoustic environments, sparse l_1 relaxations are employed to acoustically model the room only using a reduced number of microphones in [2]. However, only simple rooms (four walls, floor, and ceiling) can be modeled, and a loudspeaker emitting a known sound pattern is required. Using this technique in a previous training step, has been proved to be useful to improve source localization [32].

More recently, a novel technique for source localization in reverberant environments using wavefield sparse decomposition has been proposed in [12]. However, although it shows promising performance, the experimental results are only based on simulations and narrowband signals, which make their approach not applicable to speech signals, which is our target scenario.

3 Review of GCC based approaches

In this section we will recall the Generalized Cross Correlation and some classical procedures for source localization based on it. Some common characteristics of these algorithms are that 1) they are simple, 2) they can be used for any array topology, and 3) they don't require any extra knowledge about the signal, the noise or the room/environment. All those features make GCC based approaches to be fast, robust, and suitable for realistic scenarios.

3.1 The Generalized Cross Correlation

Let us assume that we equip an indoor environment with an array of N microphones $\mathcal{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N\}$, where $\mathbf{m}_i = (m_{ix}, m_{iy}, m_{iz})^\top$ is a three-dimensional vector denoting the position of the microphone i from a reference coordinate origin.

Given this setup, let's assume that an acoustic source is located at the generic position $\mathbf{r} = (r_x, r_y, r_z)^\top$, emitting an acoustic baseband signal $x(t)$. We denote as $x_i(t)$ the signal received by the microphone located at $\mathbf{m}_i$. The Generalized Cross Correlation (GCC) [23] obtained for each pair of microphones, m_i and m_j can be expressed as:

$$R_{i,j}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{ij}(\omega)\, X_i(\omega)\, X_j^*(\omega)\, \mathrm{e}^{\mathrm{j}\omega\tau}\, \mathrm{d}\omega \quad , \tag{1}$$

where $X_n(\omega)$ is the Fourier Transform of $x_n(t)$, and $\Psi_{ij}(\omega)$ is a weighting function that is chosen to optimize a performance criteria. A number of such functions has been studied in the literature, yielding the different variants of GCC [23].

Ideally, the τ for which $R_{i,j}(\tau)$ is maximized will correspond with the time difference of arrival (TDOA) between the two microphones. Note that the range of possible TDOA values is limited by the physical separation between the pair of microphones, $|\tau| \leq \frac{1}{c}\|\mathbf{m}_i - \mathbf{m}_j\|$, where c is the speed of the sound in air. Therefore, the TDOA estimate is calculated as

$$\hat{\tau}_{i,j} = \underset{|\tau| \leq D_{ij}}{\arg\max}\, R_{i,j}(\tau) \quad , \tag{2}$$

where $D_{i,j} = \frac{1}{c}\|\mathbf{m}_i - \mathbf{m}_j\|$ is the distance between the microphones i and j.

3.2 Time Delay Based Approaches

The former TDOA estimation is the base of the TDOA approaches. These methods are based on a two-step procedure. First, they estimate the pairwise TDOA using some variant of GCC. In a second step, the time-difference of arrival information is combined with knowledge of the microphones' positions to generate a Maximum Likelihood (ML) spatial estimator made from hyperboloids intersected in some optimal sense [6, 18].

The main problem of these methods is their robustness. In other words, since only the pairwise TDOA estimates are used in the second stage, an accurate estimation of the time delay is essential for a good performance of these methods.

Coherent noise and multi-path due to reverberation are the two major sources of error in time delay estimation, so that different approaches have been proposed to deal with them. A basic method consists in making the GCC function more robust by de-emphasizing the frequency-dependent weighting. Particularly, choosing the weighting function $\Psi_{ij}(\omega)$ as:

$$\Psi_{ij}(\omega) = \frac{1}{|X_i(\omega)\, X_j^*(\omega)|} = \frac{1}{|X_i(\omega)|\,|X_j(\omega)|} \ , \tag{3}$$

leads to the GCC-PHAT method, which has been empirically proved to reduce the influence of multipath in the GCC of a signal arriving to two different microphones in a reverberant acoustic environment. The Phase Transform (PHAT) [23, 42] has received considerable attention as the basis of speech source localization systems due to its robustness in real world scenarios [46].

Also, after the first step, the TDOA estimation can be estimated by taking advantage of the redundancy in TDOA measurements [22,34], and then proceed with the second step. For instance, in [44], the algebraic properties of a kind of matrices constructed from TDOA measurements are used in order to remove outliers in the TDOA estimations.

3.3 Steered-Beamformer-Based Locators

The previously described two-stage process requiring time-delay estimation prior to the actual local location evaluation is suboptimal. The TDOA estimation procedure represents a significant data reduction, hence it degenerates the maximum theoretical localization performance.

Beamforming based techniques [19], attempt to estimate the position of the source, by maximizing or minimizing a spatial statistic associated with each position. For instance, in the Steered Response Power (*SRP*) approach, which is the simplest beamforming method, the statistic is based on the signal power received when the microphone array is steered in the direction of a specific location. Therefore, the position of the source is supposed to be consistent with the position corresponding with the maximum estimated signal power.

Let $\mathbf{q} = (q_x, q_y, q_z)^\top$ be a generic target location at which we steer the microphone array. Then the resulting beamformed signal when the array is steered to $\mathbf{q}$ is defined as (also including the frequency domain expression):

$$y\,(t, \mathbf{q}) = \sum_{n=1}^{N} x_n\,(t + \tau_n(\mathbf{q})) \overset{\mathcal{F}}{\longleftrightarrow} Y(\omega, \mathbf{q}) = \sum_{n=1}^{N} X_n(\omega) \mathrm{e}^{\mathrm{j}\omega\tau_n(\mathbf{q})}, \tag{4}$$

where $\tau_n(\mathbf{q})$ is the propagation delay between $\mathbf{m}_n$ and $\mathbf{q}$, which is calculated as $\tau_n(\mathbf{q}) = \frac{1}{c}\|\mathbf{q} - \mathbf{m}_n\|$. The set of signals $x_1(t), \cdots, x_N(t)$ are aligned compensating the propagation delays from the target position $\mathbf{q}$ to each microphone $\mathbf{m}_n$. Usually the *delay-and-sum* beamformer, is generalized applying some adaptive filtering $H_n(\omega)$ to the signals received by the microphones, yielding

the *filter-and-sum* beamformer:

$$Y(\omega, \mathbf{q}) = \sum_{n=1}^{N} H_n(\omega)\, X_n(\omega)\, \mathrm{e}^{\mathrm{j}\omega\tau_n(\mathbf{q})} \quad . \tag{5}$$

As stated above, the *Steered Response Power* (SRP) can be expressed as the output power of the signal received from a *filter-and-sum* beamformer of N elements, but it can also be expressed in terms of the GCC [18]:

$$\begin{aligned} P(\mathbf{q}) &= \int_{-\infty}^{\infty} |Y(\omega, \mathbf{q})|^2 \, \mathrm{d}\omega = \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \int_{-\infty}^{\infty} H_i(\omega)\, H_j^*(\omega)\, X_i(\omega)\, X_j^*(\omega)\, \mathrm{e}^{\mathrm{j}\omega\Delta\tau_{ij}(\mathbf{q})}\, \mathrm{d}\omega \\ &= 2\pi \sum_{i=1}^{N}\sum_{j=1}^{N} R_{i,j}\left(\Delta\tau_{ij}(\mathbf{q})\right) \quad , \end{aligned} \tag{6}$$

were $\Delta\tau_{ij}(\mathbf{q}) = (\tau_i(\mathbf{q}) - \tau_j(\mathbf{q}))$ is the time difference of arrival between the signals arriving at microphones i and j, after being generated by a source placed at $\mathbf{q}$, and $H_i(\omega)\, H_j^*(\omega) = \Psi_{ij}(\omega)$. A particular case of SRP is when GCC-PHAT is used in equation (6) yielding *SRP-PHAT*, which is a widely used algorithm for speaker localization.

Assuming that $x_n(t)$ is a baseband signal with bandwidth ω_0 ($X_n(\omega) = 0,\ \forall\omega > \omega_0$) equation (6) can be expressed as follow:

$$P(\mathbf{q}) = 2N\omega_0 + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \int_{-\omega_0}^{\omega_0} \frac{X_i(\omega)\, X_j^*(\omega)}{|X_i(\omega)||X_j(\omega)|}\, \mathrm{e}^{\mathrm{j}\omega\Delta\tau_{ij}(\mathbf{q})}\, \mathrm{d}\omega \quad , \tag{7}$$

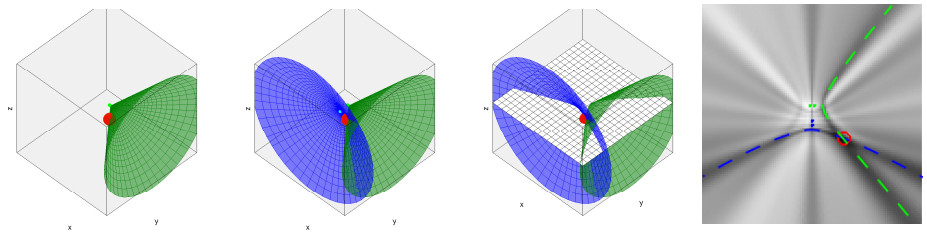
In equations (6) and (7), the terms where $i = j$ represent the power received by each single microphone. These terms have a trivial solution ($\int_{-\omega_0}^{\omega_0} \frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)||X_j(\omega)|}\mathrm{d}\omega = 2\omega_0\,, \forall i = j$) which doesn't depend on the steering position, so they only represent a known offset ($2N\omega_o$) that can be easily removed from equation (7). Without loss of generality, we will not take this offset term into account hereinafter (this will imply the appearance of negative values in $P(\mathbf{q})$).

To ease further mathematical development, we group the microphones in different pairs, described as elements in a set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_P}\}$, where $\mathbf{p}_j = \{\mathbf{m}_{j_1}, \mathbf{m}_{j_2}\}$ is composed of two three-dimensional vectors, $\mathbf{m}_{j_1} \in \mathcal{M}$ and $\mathbf{m}_{j_2} \in \mathcal{M}$, with $\mathbf{m}_{j_1} \neq \mathbf{m}_{j_2}$, describing the spatial location of the microphones in pair j. If all microphone pairs are allowed then $N_P = N(N-1)/2$.

Equation (7) can be rewritten taking into account the contributions of each pair of microphones:

$$\begin{aligned} P\left(\mathbf{q}\right) &= \sum_{j=1}^{N_P} \left[\int_{-\omega_0}^{\omega_0} 2\frac{\left(X_{j_1}(\omega)\, X_{j_2}^*(\omega)\mathrm{e}^{\mathrm{j}\omega\Delta\tau(\mathbf{p}_j,\mathbf{q})}\right)}{|X_{j_1}(\omega)||X_{j_2}(\omega)|}\, \mathrm{d}\omega \right] = \\ &= 4\pi \sum_{j=1}^{N_P} R_{j_1,j_2}\left(\Delta\tau(\mathbf{p}_j,\mathbf{q})\right) = 4\pi \sum_{j=1}^{N_P} R_j\left(\Delta\tau(\mathbf{p}_j,\mathbf{q})\right) \quad , \end{aligned} \tag{8}$$

where $\Delta\tau(\mathbf{p}_j, \mathbf{q}) = (\tau_{j_1}(\mathbf{q}) - \tau_{j_2}(\mathbf{q}))$ is the difference in arrival times of the acoustic signal to reach the microphones in pair $\mathbf{p}_j$ ($\mathbf{m}_{j_1}$ and $\mathbf{m}_{j_2}$), that is, the required delay to steer the microphone



(a) Hyperboloid with equal TDOA for a given pair of microphone.

(b) Intersection of two hyperboloids for two pairs of microphones.

(c) Intersection of the hyperboloids cut by a plane

(d) Real *SRP-PHAT* map with hyperbolas superimposed

Figure 1: Geometric interpretation of GCC based approaches

pair $\mathbf{p}_j$ to the location $\mathbf{q}$. Hereafter $R_j = R_{j_1,j_2}$ will be employed to refer to the GCC-PHAT of the pair of microphones $\mathbf{p}_j$.

SRP-PHAT is usually defined as a reference standard for source localization, because of its simplicity and robustness in reverberant and noisy environments, being a widely used algorithm for speaker localization [3, 8, 15, 20, 21, 43]

The main drawback of beamforming methods is their high computational cost, provided that they sample all potential positions of the space, labeling all local maxima as position candidates for acoustic sources. Hence there is a trade-off between the computational cost and the number of sampled positions, Q, which is proportional to the desired resolution and the dimension of the explored region. The grid, $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_Q\}$, is defined as the set of Q sampled positions.

4 Redundancy of PHAT based approaches

As stated above, PHAT is one of the preferred techniques due to its excellent performance in low noise, reverberant environments [46]. Nevertheless, in this section we will show that PHAT based approaches are suboptimal since they produce quite redundant results. There are two kinds of redundancy: 1) Geometric, which is inherent to how SRP is formed and doesn't provide extra information and hence it should be removed; 2) temporal redundancy, which has been deeply studied in [42] and can be used to increase the robustness in localization as we will show in section 6.

4.1 Geometric redundancy

Assume that only one microphone pair, e.g. pair $\mathbf{p}_j$, is placed in the environment, thus the *SRP-PHAT* power estimation at a generic position, $\mathbf{r}$, can be calculated as $P(\mathbf{r}) = 4\pi R_j(\Delta\tau(\mathbf{p}_j, \mathbf{r}))$

If we define $\mathbf{q}_r$ as any location for which $\Delta\tau(\mathbf{p}_j, \mathbf{q}_r) = \Delta\tau(\mathbf{p}_j, \mathbf{r})$, the corresponding cross-correlation values $R_j(\Delta\tau(\mathbf{p}_j, \mathbf{q}_r))$ will be identical to $R_j(\Delta\tau(\mathbf{p}_j, \mathbf{r}))$. For a microphone pair, it can be easily demonstrated [1] that the geometric place of points $\mathbf{q}_r$ is one of the sheets of a two-sheeted hyperboloid of revolution, whose foci are located at the microphone locations, as shown in Fig. 1a. Since $R_j(\tau)$ is a good approximation of the likelihood function of TDOA between two microphones [23, 46], its hyperbolic spread version $R_j(\Delta\tau(\mathbf{p}_j, \mathbf{r}))$ can be regarded as the spatial likelihood function for the position of the source.

Additional pairs will generate new spatial likelihood functions and it will allow us to disambiguate the actual position of the acoustic source. Time Delay Based Approaches intersect the maximum likelihood hyperboloid calculated for each pair of microphones [11] (figure 1b).

On the other hand, SRP-PHAT combines all those functions yielding a global spatial likelihood function which is more robust. Nevertheless, hyperbolic spreading of GCC functions can be still appreciated in SRP-PHAT maps, especially when only a few pairs of microphone are used. In Fig. 1d the hyperbolas corresponding to the maximum likelihood TDOA estimator for each microphone pair have been superimposed to the SRP-PHAT map which has been calculated at a plane located 61cm above the microphone locations as shown in figure 1c. When the number of microphone pairs used is high, it becomes harder to identify each pair contribution within the SRP-PHAT map.

Hence, the first conclusion is that while SRP-PHAT combines all the corresponding GCC-PHAT in a robust manner, it is very redundant. Besides, the contribution of each pair of microphones is unrecoverable from SRP-PHAT making difficult solving ambiguities. Since SRP-PHAT is constructed using equation (8) , it is evident that the pairwise GCC-PHAT contains equivalent information to SRP-PHAT. Hence the vector $\phi = [R_1(t_1), R_2(t_2), \cdots, R_{N_P}(t_{N_P})]^T$, which concatenates all the correlations, is usually is a more compact way to represent the information available in $P(\mathbf{q})$ and also keeps clear the contribution of each microphone pair. Moreover, the number of required samples in $R_j(t_j)$ only depends on the distance between microphones ($|t_j| < \frac{1}{c}\|\mathbf{m}_{j_1} - \mathbf{m}_{j_2}\|$), and the sampling frequency, not on the size of the grid.

4.2 Temporal Redundancy

In [42] we derive an analytic model for accurately predicting the behavior of the GCC-PHAT correlation for wideband signals, taking into account both the room geometry and the microphone array topology. We also show that such model is independent of the spectral content of the recorded signals, for both anechoic and reverberant conditions.

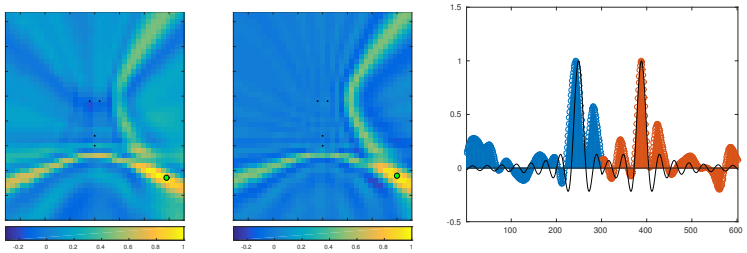
For simplicity, at this point, let us consider the anechoic propagation scenario in [42] wherein a single source generates a baseband signal with bandwidth ω_0. In that case, the signal at microphone n can be represented as a time-shifted and attenuated version of $X(\omega)$, i.e. $X_n(\omega) = \alpha_n X(\omega) \mathrm{e}^{-\mathrm{j}\omega\tau_n(\mathbf{r})}$ where α_n is a microphone dependent attenuation factor, and $\tau_n(\mathbf{r})$ is the propagation delay between $\mathbf{m}_n$ and $\mathbf{r}$, which is calculated as $\tau_n(\mathbf{r}) = \frac{1}{c}\|\mathbf{r} - \mathbf{m}_n\|$.

In the former scenario, the resulting GCC-PHAT described by equations (1) and (3) for the pair $\mathbf{p}_j$ can be approximated as a sinc function ($\mathrm{sinc}(x) = \frac{\sin(x)}{x}$), as shown below [42]:

$$\begin{aligned} \bar{R}_j(t_j, \mathbf{r}) &= \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} \frac{X_{j_1}(\omega)\, X_{j_2}^*(\omega)}{|X_{j_1}(\omega)|\,|X_{j_2}(\omega)|}\, \mathrm{e}^{\mathrm{j}\omega t_j}\, \mathrm{d}\omega \approx \\ &= \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} \mathrm{e}^{-\mathrm{j}\omega\left(\tau_{j_1}(\mathbf{r}) - \tau_{j_2}(\mathbf{r})\right)}\, \mathrm{e}^{\mathrm{j}\omega t_j}\, \mathrm{d}\omega = \\ &= \frac{\omega_0}{\pi}\mathrm{sinc}\left(\omega_0\left(t_j - \Delta\tau_j(\mathbf{r})\right)\right) \quad , \end{aligned} \tag{9}$$

where $\Delta\tau_j(\mathbf{r}) = (\tau_{j_1}(\mathbf{r}) - \tau_{j_2}(\mathbf{r}))$ is the time-difference-of-arrival (TDOA) between the microphones j_1 and j_2. It is noteworthy that, for any two microphones, only a small set of TDOAs are physically feasible as $|\Delta\tau_j(\mathbf{r})| \leq \frac{1}{c}\|\mathbf{m}_{j_1} - \mathbf{m}_{j_2}\|$. This model has also been successfully employed to derive a new technique of calibration in diffuse noise [41].

In the next sections we describe additional effects that can be taken into account: discretization and reverberation.



(a) Measured SRP (b) Pattern (c) Correlations

Figure 2: Example

4.2.1 Discretization effects

So far, continuous time signals have been considered. Nevertheless, in real systems, discretization and windowing of the physical signal are mandatory. Assuming the Nyquist criterion is fulfilled and a that the analysis window size is large enough, discretization effects lead the following expression [42]:

$$\begin{aligned}\bar{R}_j(k_j, \mathbf{r}) &= \frac{1}{N_F} \sum_{l=-M_0}^{M_0} \mathrm{e}^{-\mathrm{j}\frac{2\pi\Delta\tau_j(\mathbf{r})}{N_F T_s}} \mathrm{e}^{\mathrm{j}\frac{2\pi k_j}{N_F}} \\ &= \frac{1}{N_F} \frac{\sin\left(\frac{2\pi(M_0+0.5)}{N_F T_s}\left(k_j T_s - \Delta\tau_j(\mathbf{r})\right)\right)}{\sin\left(\frac{\pi}{N_F T_s}\left(k_j T_s - \Delta\tau_j(\mathbf{r})\right)\right)} \quad ,\end{aligned} \tag{10}$$

where N_F is the number of samples of the window after zero padding and $T_s = \frac{1}{f_s}$ is the sampling period, $k_j = t_j/T_s$ and M_0 is the DFT index corresponding to the signal bandwidth ω_0:

$$M_0 = \left\lfloor \frac{\omega_0 T_s N_F}{2\pi} \right\rfloor = \left\lfloor \frac{f_0 N_F}{f_s} \right\rfloor \quad . \tag{11}$$

4.2.2 Reverberation

Reverberation effects can be also modeled if the room geometry is known. Nevertheless, since this knowledge is not always available, we have not considered it in this work. The interested reader is encouraged to refer to [42], where this and other effects have been exhaustively described.

4.2.3 Examples

In Fig. 2a we show the SRP-PHAT map measured in a slice located at the speaker's mouth height for a frame where only one speaker was active (the position of the speaker has been marked with a green dot). Analyzing this figure, it can be clearly seen how the acoustic energy produced by the speaker is spread over the space, making it difficult to accurately localize the speaker. Nevertheless, such behavior is not at random, but can be predicted combining equations (8) and (10). As an example, the predicted SRP for a speaker in some position is shown Fig. 2b, the similarities between both, the measured and predicted SRP-PHAT map are more than evident.

The temporal redundancy is even more evident in the vector ϕ. As an example, in figure 2c the vector $\phi = [R_1(t_1), R_2(t_2)]^T$ corresponding to figure 2a is shown, where the samples belonging

to $R_1(t_1)$ and $R_2(t_2)$ have been plotted in different color, blue and orange respectively. Besides, the generative model for the ground truth created from equation (10) has been superimposed in a black solid line.

Figure 2c makes evident that the model correctly predicts behaviors due to the direct path propagation. According to [42], the rest of the artifacts are caused by the multipath and can be modeled as a linear combination of sinc functions. Nevertheless, as stated before, multipath effects are hard to predict (even more when the geometry of the room is unknown) but can be discarded using the pairwise redundancy [44].

5 Multiple Sources

In this section we will study the multi-source case, which is quite more intricate due the cross-correlation terms between each sources. Let us assume, without loss of generality, that s sources located at the positions $\mathbf{r}_1, \cdots, \mathbf{r}_s$ are active at the same time. The signal at the microphone n is then the sum of the contribution from each source, i.e. $X_n(\omega) = \sum_{a=1}^{s} X_{a,n}(\omega)$ where $X_{a,n}(\omega)$ is the Fourier Transform of the contribution of the source a at the microphone n.

Thus, the GCC-PHAT obtained in a multisource scenario has the following expression:

$$\bar{R}_j(t_j, \mathbf{r}_1, \cdots, \mathbf{r_s}) = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} \frac{\sum_{a=1}^{s}\sum_{b=1}^{s} X_{a,j_1}(\omega) X^*_{b,j_2}(\omega)\, \mathrm{e}^{\mathrm{j}\omega t_j}}{\left|\sum_{a=1}^{s} X_{a,j_1}(\omega)\right| \left|\sum_{b=1}^{s} X_{b,j_2}(\omega)\right|} \mathrm{d}\omega \ . \tag{12}$$

Assuming anechoic propagation for each source, i.e. $X_{a,n}(\omega) = X_a(\omega)\mathrm{e}^{-\mathrm{j}\omega\tau_n(\mathbf{r}_a)}$ where $X_a(\omega)$ is the signal emitted by the source a, the numerator of equation (12) can be rewritten as:

$$\sum_{a=1}^{s}\sum_{b=1}^{s} X_{a,j_1}(\omega) X^*_{b,j_2}(\omega)\, \mathrm{e}^{\mathrm{j}\omega t_j} = \sum_{a=1}^{s} \left(|X_a(\omega)|^2 + \frac{1}{2}\sum_{b\neq a} C_{j,a,b}(\omega) \right) \mathrm{e}^{-\mathrm{j}\omega(t_j - \Delta\tau_j(\mathbf{r}_a))} =$$

$$= \sum_{a=1}^{s} \Psi_{j,a}(\omega) \mathrm{e}^{-\mathrm{j}\omega(t_j - \Delta\tau_j(\mathbf{r}_a))} \ , \tag{13}$$

where $C_{j,a,b}(\omega) = X_a(\omega)X^*_b(\omega)\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_1}(\mathbf{r}_a) - \tau_{j_1}(\mathbf{r}_b))} + X_b(\omega)X^*_a(\omega)\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_2}(\mathbf{r}_b) - \tau_{j_2}(\mathbf{r}_a))}$ is a term related to the cross-correlation between $X_a(\omega)$ and $X_b(\omega)$.

On the other hand, the denominator of equation (12) can be rewritten as:

$$\left|\sum_{a=1}^{s} X_{a,j_1}(\omega)\right| \left|\sum_{b=1}^{s} X_{b,j_2}(\omega)\right| = \prod_{i=1}^{2} \left(\sum_{a=1}^{s} \left\{ |X_a(\omega)|^2 + \sum_{b\neq a} \mathrm{Re}\left(X_a(\omega)X^*_b(\omega)\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_i}(\mathbf{r}_a) - \tau_{j_i}(\mathbf{r}_b))} \right) \right\} \right)^{\frac{1}{2}} . \tag{14}$$

When the coherence [10] between each source is small, i.e:

$$\sum_{a=1}^{s} |X_a(\omega)|^2 > \sum_{a=1}^{s}\sum_{b\neq a} \mathrm{Re}\left(X_a(\omega)X^*_b(\omega)\mathrm{e}^{-\mathrm{j}\omega(\tau_{j_i}(\mathbf{r}_a) - \tau_{j_i}(\mathbf{r}_b))} \right) \ , \tag{15}$$

and applying first-degree taylor polynomial of square root ($\sqrt{1+x} \approx 1 + \frac{x}{2}$), equation (14) can be approximated to:

$$\sum_{a=1}^{s} \left\{ |X_a(\omega)|^2 + \frac{1}{2}\sum_{b\neq a}^{s} \mathrm{Re}\left(C_{j,a,b}(\omega)\right) \right\} = \sum_{a=1}^{s} \mathrm{Re}\left(\Psi_{j,a}(\omega)\right) \ , \tag{16}$$

Therefore, from equations (13) and (16),we can rewrite equation (12) as:

$$\bar{R}_j(t_j, \mathbf{r}_1, \cdots, \mathbf{r_s}) \approx \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} \frac{\sum_{a=1}^{s} \Psi_{j,a}(\omega) \mathrm{e}^{-\mathrm{j}\omega(t_j - \Delta\tau_j(\mathbf{r}_a))}}{\sum_{a=1}^{s} \mathrm{Re}\left(\Psi_{j,a}(\omega)\right)} \mathrm{d}\omega \quad . \tag{17}$$

Finally, if all the interfering signals have similar spectral density, then equation (17) becomes independent of the spectral content of the sources. Thus, GCC-PHAT can be modeled as the sum of s sinc functions:

$$\bar{R}_j(t_j, \mathbf{r}_1, \cdots, \mathbf{r_s}) \approx \sum_{a=1}^{s} \beta_a \,\mathrm{sinc}\left(\omega_0 \left(t_j - \Delta\tau_j(\mathbf{r}_a)\right)\right) \quad , \tag{18}$$

where β_a is a coefficient that weight the contribution of each source.

It is important to note that the more sources in the scene, the more restrictive equation (15) becomes. Nevertheless, the former condition is always satisfied for less than three sources.

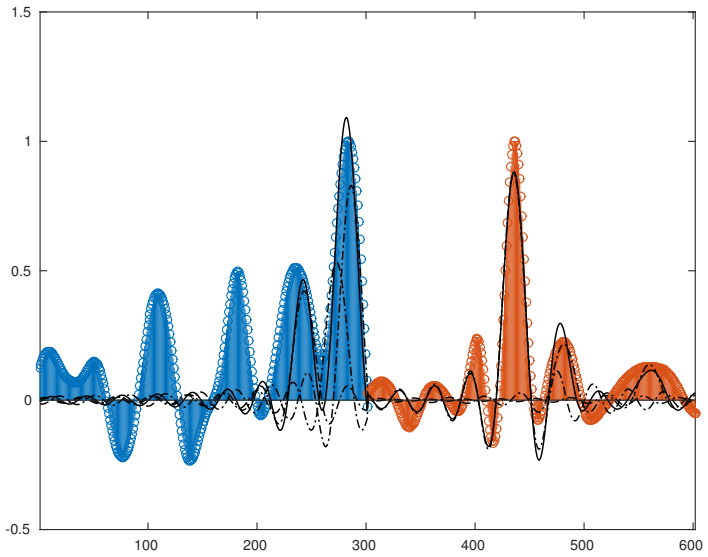


Figure 3: Example of a correlation

In Figure 3 we show an example of two obtained correlations for multiple (three) speakers. The sample of each correlation have been represented in different colors (blue and orange). The generative model for the ground truth and the contribution of each speaker has been superposed in solid and dotted black line respectively.

As in figure 2c, some other peaks appear due multipath, moreover, since the number of sources has increased the number of reflections is also bigger. Thus, more redundancy (i.e, more pairs of microphones) is needed to determine correctly the position of the sources in a multisource scenario.

6 Sparse Localization

As stated in section 3.3 using SRP-PHAT has been historically useful since it has an easy interpretation. Nevertheless SRP-PHAT has also some drawbacks that we have discussed along

section 4, 1) it doesn't take into account temporal redundancy, 2) it is not possible to recover the contribution from each pair of microphones, 3) the number of measurements, Q, is proportional to the desired resolution and the dimension of the explored region.

As consequence of the first issue point, some uncertainty may appear, for example one might believe that another speaker is present at the bottom left corner in Fig. 2a (corresponding to the high values of the SRP-PHAT function in the lower hyperbola).

The two last weaknesses of SRP can be solved using the previously introduced vector ϕ which concatenates all the correlations. As stated above, it is evident from equation (8) that ϕ contains information that is equivalent to that contained in SRP-PHAT, while removing spatial redundancy. Note that, unlike SRP, the number of samples in ϕ does not depends on the desired spatial resolution (i.e. Q), but it relies on the sampling rate and the number of microphone pairs.

6.1 Problem statement

The aim of this work is to propose a new localization algorithm which uses temporal redundancy for improving accuracy, keeping the number of microphones low. To do this, we will fit the acquired data with a model generated from equation (10). If all the sources satisfy the properties discussed in section 5, then we have seen that the correlation is approximately linear. Therefore, it yields the following non-convex optimization.

$$\underset{\mathbf{a}_1,\cdots,\mathbf{a_k},\mathbf{q}_1,\cdots,\mathbf{q_k}}{\text{minimize}} \quad \left\| \phi - \sum_{i=1}^{k} a_i \, \bar{\phi}(\mathbf{q}_i) \right\|^2 \quad . \tag{19}$$

where $\bar{\phi}(\mathbf{q}_i) = [\bar{R}_1(t_1, \mathbf{q}_i), \bar{R}_2(t_2, \mathbf{q}_i), \cdots, \bar{R}_{N_P}(t_{N_P}, \mathbf{q}_i)]^T$ is the predicted response for a source located at position $\mathbf{q}_i$.

Assuming that the position of the sources are constrained to a finite set $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_Q\}$ of Q positions, the former problem can be reformulated as:

$$\min_{\mathbf{a}} \|\phi - \Phi \, \mathbf{a}\|^2 \quad \text{s.t.} \quad \|\mathbf{a}\|_0 \leq k \ , \tag{20}$$

where k ($k \leq Q$) is the maximum number of speakers and $\Phi = [\bar{\phi}(\mathbf{q}_1), \cdots, \bar{\phi}(\mathbf{q}_Q)]$ is a matrix which contains the models for each of the possible locations.

Consequently, the support of the optimal $\mathbf{a}$, $\mathbf{a}^*$, is related to the estimated position of the sources. Thus, if $i \in \text{supp}(\mathbf{a}^*)$ then, our method estimates that a source is placed at $\mathbf{q}_i$.

The problem described by equation (20) is hard to solve since it is NP-Hard and non-convex. Despite its theoretical complexity, several methods and approximations have been proposed so far, and of special relevance are those methods based on using the l_1 norm as a convex relaxation of the l_0 norm [39, 40]. This relaxation transforms equation (20) into the following:

$$\min_{\mathbf{x}} \|\phi - \mathbf{\Phi a}\|^2 + \lambda \|\mathbf{a}\|_1 \tag{21}$$

where λ is the Lagrange multiplier and has a direct relationship with k. Problem (21) is convex, thus convergence is guaranteed and can be solved in polynomial time.

The problem of finding a least squares estimation subject to a l_1 restriction has been independently presented and popularized under the names of *Least Absolute Shrinkage Selection Operator (LASSO)* [37] and *Basis Pursuit Denoising* [13], being object of intensive study. In the past few years numerous optimization methods have been proposed, some of them adapted to specific problems.

Solving the relaxed problem (21) does not necessary imply finding the solution to the original l_0 problem. The closeness and validity of l_1 relaxations has been extensively studied [39]. In

some problems, the structure of matrix $\mathbf{\Phi}$ and the expected degree of sparsity in the solution can make l_1 relaxations to be exact. For general linear systems, as it is the case in this paper, where matrix $\mathbf{\Phi}$ has no apparent structure, l_1 relaxation empirically tends to impose only approximate sparse solutions.

6.2 Single source case

When we previously know that only one source is active, i.e $k = 1$, the problem described in equation (19) can be drastically simplified:

$$\min_{a,\mathbf{q}} \left\|\phi - a\,\bar{\phi}(\mathbf{q})\right\|^2 = \min_{a,\mathbf{q}} \|\phi\|^2 - 2a\bar{\phi}(\mathbf{q})^\top\phi + a^2\left\|\bar{\phi}(\mathbf{q})\right\|^2 \quad . \tag{22}$$

In order to find the solution of equation (22), first we will look for the optimal value of a:

$$a^* = \frac{\bar{\phi}(\mathbf{q})^\top\phi}{\left\|\bar{\phi}(\mathbf{q})\right\|^2} \quad , \tag{23}$$

and then, replacing it again in equation (22), we will obtain a new formulation of the problem:

$$\min_{\mathbf{q}} \|\phi\|^2 - \frac{\left(\bar{\phi}(\mathbf{q})^\top\phi\right)^2}{\left\|\bar{\phi}(\mathbf{q})\right\|^2} = \max_{\mathbf{q}} \frac{\left(\bar{\phi}(\mathbf{q})^\top\phi\right)^2}{\left\|\bar{\phi}(\mathbf{q})\right\|^2} \quad , \tag{24}$$

which only depends on the position of the source, $\mathbf{q}$.

Note that the problem described by equation (24) is non-convex. Nevertheless, we can get a convex approximation constraining the position of the source, $\mathbf{q}$, to a finite set $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_Q\}$ of Q positions:

$$\max_{\mathbf{q}} \frac{\left(\bar{\phi}(\mathbf{q})^\top\phi\right)^2}{\left\|\bar{\phi}(\mathbf{q})\right\|^2} \quad \text{s.t.} \quad \mathbf{q} \in \mathcal{Q} \quad . \tag{25}$$

Consequently, the localization problem can be interpreted as a simple atom selection problem: we are looking for the atom $\bar{\phi}(\mathbf{q}_i)$, $\mathbf{q}_i \in \mathcal{Q}$ which better fits the measured data, ϕ. It is equivalent to solve the following problem:

$$\max_{\mathbf{i}} \left|\hat{\phi}_i^\top\phi\right| \quad , \tag{26}$$

where $\hat{\phi}_i = \bar{\phi}(\mathbf{q}_i)/\|\bar{\phi}(\mathbf{q}_i)\|$ is the unitary vector with the same direction as the atom $\bar{\phi}(\mathbf{q}_i)$. Problem (26) has a straightforward solution, which allows an efficient solution.

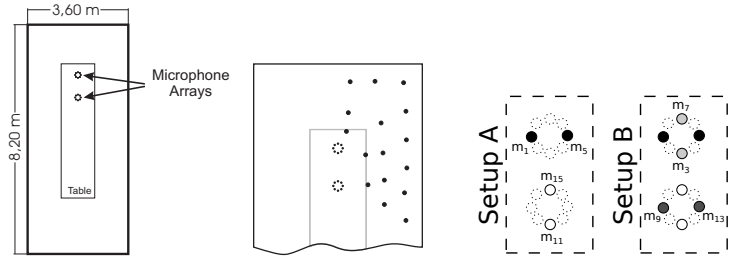
As we will see later on, the proposed algorithm for single source is faster than SRP-PHAT and achieve better performance in localization.

6.3 Solution refinement

In order to make the localization problem convex, we had to discretize the search space. Convexity is a very convenient property since it avoids the convergence of the algorithms in local minimums, and provides efficient solutions. On the other hand, a non-convex optimization problem as equation (19) will reach the global optimal solution only if it is correctly initialized.

It seems a good idea using the solution of convex problem (21) (the equation (26) in single source scenario) as the initialization of the non-convex problem (19). As we will see in the next section, by doing this we have improved the localization performance in the single source scenario.

As future work, we will thoughroughly evaluate this method in a multi-source scenario, and we will perform an in-depth study about the maximum resolution an limits of this method.



(a) IDIAP room layout showing the centered table, and the microphones arranged in two circular arrays.

(b) Positions evaluated in the real data experiments. Only the relevant room section is shown.

(c) Microphone pairs setups used in the example

Figure 4: IDIAP Smart Meeting Room: experimental details.

7 Experiments and discussion

In this section, we will present some preliminary experimental results obtained in a real environment using the AV16.3 database recordings [25].

7.1 Experimental setup

The AV16.3 database is an audio-visual corpus recorded in the *Smart Meeting Room* of the IDIAP research institute, in Switzerland. The IDIAP Meeting Room consists on a 8.2mx3.6mx2.4m rectangular room containing a centrally located 4.8mx1.2m rectangular table, on top of which two circular microphone arrays of radius 0.1 meters are located, each of them composed by 8 microphones. The centers of the two arrays are separated by 0.8m and the origin of coordinates is located in the middle point between the two arrays.

For the single source scenario, we are using only 4 or 8 microphones (out of the 16 available in the AV16.3 data set), grouped in two or four microphone pairs. The selected microphone pairs configurations are shown in Figure 4c, in which microphones with the same color are considered as belonging to the same microphone pair.

Possible speakers' locations are distributed along a L-shaped area around the table as seen in Figures 4a and 4b. A detailed description of the meeting room, can be found in [28].

For multisource scenario, more microphones pairs are needed in order to avoid uncertainties. Thus we have made use of all available pair of microphones.

The audio recordings are synchronously sampled at 16 KHz, and the complete database along with the corresponding annotation files containing the recordings ground truth is fully accessible on-line at [24]. It is composed by several sequences or recordings with varying number of speakers involved and their activity.

7.2 Single Source Results

In this experiment we have employed the sequence *seq01-1p-0000* in the aforementioned database. In this sequence, since the height of the source's is constant (0,61 above the microphones on the table) we can restrict the experiment to a 2D scenario. Consequently, the grid $\mathcal{Q}$ is composed of the uniformly sampled locations in an regular grid contained in a plane 61 cm above the

Table 1: Results for SETUP A

	SRP	Projection	Projection + refinement
Average Localization error (mm)	1010	1034	988
Average Fine Localization error (mm)	292	220	217
Pcor(%)	42.82	39.29	40.69

Table 2: Results for SETUP B

	SRP	Projection	Projection + refinement
Average Localization error (mm)	767	727	695
Average Fine Localization error (mm)	287	200	209
Pcor(%)	59.15	57.10	59.23

microphone arrays. The resolution of the grid (i.e the minimum distance between two locations) was 10cm.

The sequence duration is 208 seconds in total, which has been divided in 320 ms frames with a frame shift of 80ms. The total number of frames labeled was 1219.

In table 1, we show the results obtained for the method described in section 6.2 using the setup-A (two pair of microphones). The first column refers to the results obtained with SRP-PHAT. On the other hand, the results obtained after applying the algorithm described in section 6.2 are displayed in the second column of the table ("Projection"). The 'Average Fine Localization error' refers to the average localization error of those estimation with localization error less than 0.5 m. Finally PCor is the percentage rate of fine estimations.

The proposed algorithm performs very similar to SRP-PHAT. Nevertheless it obtains a significant error reduction in the fine error, keeping a similar Pcor.

It is also important to note that the results obtained by simple projection can be improved via the refinement step described by equation (19) (particularizing $k = 1$). These results are shown in the last column of the table ("Projection +refinement").

The results obtained using the Setup-B (4 pairs of microphones) are consistent with the previous results, as shown in table 2.

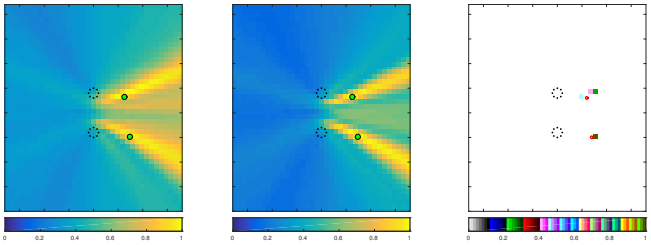
7.3 Multi-source results

As we have previously seen, the multi-source case is quite more intricate than single source case. Under the assumptions considered in section 5, we can assume than the total response is the sum of the responses for each source.

In this section we will show some representative frames of the sequence *seq37-3p-0001* in AV16.3 database. The aim is to provide a glance of the behavior of the algorithm described in a multisource scenario.

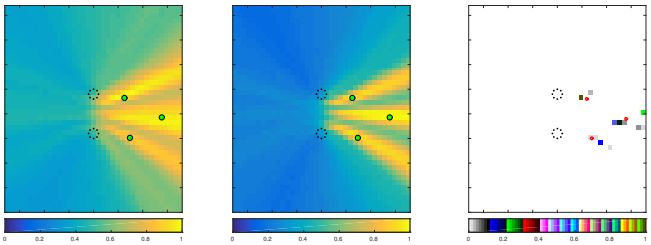
In figures 5 and 6 we show the solution of the problem (21) when two and three speakers are active respectively. In all image microphones have been represented as black dots (shaped in circles in the center of the images).

In both cases, the left figure is the obtained SRP-PHAT where the real positions of the speakers have been represented with a green dot. Although we can distinguish a beam for each source, the position of the sources is unclear since the maximum of the beam is spread.



(a) Measured SRP (b) Pattern (c) Sparse solution

Figure 5: Two sources example



(a) Measured SRP (b) Pattern (c) Sparse solution

Figure 6: Three source example

On the other hand, in the right side figures we represent the projection of the solution of equation (21) on the xy plane. Such solution is much more sparse than SRP-PHAT and moreover all the non-zero elements (colored squares) are concentrated around the speakers' locations (red circles).

Since for these examples 120 different correlations have been used, it is not possible to represent all the correlations in a single figure. Thus we consider more informative to show the SRP-PHAT generated from the proposed model (images in the middle). The reader can verify the similarities between the proposed model an the measured SRP-PHAT.

The results obtained are still preliminary. Generating metrics and compare them with other methods is still a pending work.

8 Conclusions

In this paper, we have proposed a novel method to localize active acoustic sources using sparse constraints. We have extended the generative model proposed in [42] to multiple sources, and demonstrated that under reasonable conditions, small coherence between sources and similar spectral density, GCC-PHAT has a linear behaviour. According to the former point, we propose a linear generative model for the measured GCC-PHAT functions for several pair of microphones.

Localization is performed via regression analysis with sparse constraints, assuming that only few sources are active at the same time. We also propose a convex relaxation of the localization problem in order to make it computationally tractable.

Finally, we have performed some experiments with real data in single and multi-source scenario getting preliminary but promising results.

References

[1] Xavier Alameda-Pineda and Radu Horaud. A geometric approach to sound source localization from time-delay estimates. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(6):1082–1095, 2014.

[2] D. Ba, F. Ribeiro, Cha Zhang, and D. Florêncio. L1 regularized room modeling with compact microphone arrays. In *Proceedings of ICASSP 2010*, pages 157 –160, march 2010.

[3] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi. Evaluating real-time audio localization algorithms for artificial audition in robotics. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2033 –2038, oct. 2009.

[4] R.G. Baraniuk. Compressive sensing [lecture notes]. *Signal Processing Magazine, IEEE*, 24(4):118–121, 2007.

[5] Michael Brandstein and Darren Ward, editors. *Microphone Arrays : Signal Processing Techniques and Applications*. Springer, 2001.

[6] Michael S. Brandstein and Harvey F. Silverman. A practical methodology for speech source localization with microphonearrays. *Computer Speech & Language*, 11(2):91–126, 1997.

[7] H. Buchner, R. Aichner, and W. Kellermann. *Blind Speech Separation*, chapter TRINICON-based blind system identification with application to multiplesource localization and separation, pages 101–147. Springer-Verlag, September 2007.

[8] Taras Butko, Fran Gonzalez Pla, Carlos Segura Perales, Climent Nadeu Camprubí, and Francisco Javier Hernando Pericás. Two-source acoustic event detection and localization: online implementation in a smart-room. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO'11)*, pages 1317–1321, 2011.

[9] E.J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[10] G Clifford Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, 1987.

[11] YT Chan and KC Ho. A simple and efficient estimator for hyperbolic location. *Signal Processing, IEEE Transactions on*, 42(8):1905–1915, 1994.

[12] G. Chardon and L. Daudet. Narrowband source localization in an unknown reverberant environment using wavefield sparse decomposition. In *Proceedings of ICASSP 2012*, 2012.

[13] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, pages 129–159, 2001.

[14] J.F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826, 1973.

[15] M. Cobos, A. Marti, and J.J. Lopez. A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. *Signal Processing Letters, IEEE*, 18(1):71–74, 2011.

[16] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[17] J.H. DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, 2000.

[18] JosephH. DiBiase, HarveyF. Silverman, and MichaelS. Brandstein. Robust localization in reverberant rooms. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays*, Digital Signal Processing, pages 157–180. Springer Berlin Heidelberg, 2001.

[19] Jacek P. Dmochowski and Jacob Benesty. Steered beamforming approaches for acoustic source localization. In Israel Cohen, Jacob Benesty, and Sharon Gannot, editors, *Speech Processing in Modern Communication*, volume 3 of *Springer Topics in Signal Processing*, pages 307–337. Springer Berlin Heidelberg, 2010.

[20] J.P. Dmochowski, J. Benesty, and S. Affes. A generalized steered response power method for computationally viable source localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2510 –2526, nov. 2007.

[21] Hoang Do and H.F. Silverman. Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data. In *Proceedings of ICASSP 2010*, pages 125 –128, march 2010.

[22] W. Hahn and S. Tretter. Optimum processing for delay-vector estimation in passive signal arrays. *Information Theory, IEEE Transactions on*, 19(5):608–614, Sep 1973.

[23] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320 – 327, aug 1976.

[24] Guillaume Lathoud. Av16.3 dataset. http://www.idiap.ch/dataset/av16-3/, [Last accessed in july 2015].

[25] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. AV16.3: An audio-visual corpus for speaker localization and tracking. In Samy Bengio and Hervé Bourlard, editors, *Procceedings of the MLMI*, volume 3361 of *Lecture Notes in Computer Science*, pages 182–195. Springer-Verlag, 2004.

[26] D. Malioutov, M. Cetin, and A.S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. In *IEEE Transactions on Signal Processing* [27], pages 3010 – 3022.

[27] D.M. Malioutov. *A Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays*. PhD thesis, Massachusetts Institute of Technology, 2003.

[28] Darren C. Moore. The idiap smart meeting room. Technical report, IDIAP Research Institute, Switzerland, November 2004.

[29] M. Omologo and P. Svaizer. Use of the cross-power-spectrum phase in acoustic event location. *IEEE Trans. on Speech and Audio Processing*, 5:288–292, 1993.

[30] P. Pertilä and M.S. Hämäläinen. A track before detect approach for sequential bayesian tracking of multiple speech sources. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing*, pages 4974 – 4977, 2010.

[31] B.D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Transactions on Signal Processing*, 47(1):187–200, 1999.

[32] F. Ribeiro, Demba Ba, Cha Zhang, and D. Floêncio. Turning enemies into friends: Using reflections to improve sound source localization. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 731 –736, july 2010.

[33] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, 4(13):593–606, July 2005.

[34] Hing Cheung So, Yiu Tong Chan, and F.K.W. Chan. Closed-form formulae for time-difference-of-arrival estimation. *Signal Processing, IEEE Transactions on*, 56(6):2614–2620, June 2008.

[35] Ke Sun, Yimin Liu, Huadong Meng, and Xiqin Wang. Adaptive sparse representation for source localization with gain/phase errors. *Sensors*, 11(5):4780–4793, 2011.

[36] V.N. Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3(1):33–107, 2003.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[38] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[39] J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[40] Joel A. Tropp. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing*, 86(3):589 – 602, 2006.

[41] J. Velasco, M.J. Taghizadeh, A. Asaei, H. Bourlard, C.J. Martin-Arguedas, J. Macias-Guarasa, and D. Pizarro. Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2669–2673, April 2015.

[42] Jose Velasco, Carlos J. Martín-Arguedas, Javier Macias-Guarasa, Daniel Pizarro, and Manuel Mazo. Proposal and validation of an analytical generative model of srp-phat power maps in reverberant scenarios. *Signal Processing*, 119:209 – 228, feb 2016.

[43] Jose Velasco, Daniel Pizarro, and Javier Macias-Guarasa. Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints. *Sensors*, 12(12):13781–13812, oct 2012.

[44] Jose Velasco, Daniel Pizarro, Javier Macias-Guarasa, and Afsaneh Asaei. Tdoa matrices: Algebraic properties and their application to robust denoising with missing data. *IEEE Transactions on Signal Processing*, 2016. Under revision.

[45] Alexander Waibel and Rainer Stiefelhagen. *Computers in the Human Interaction Loop*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[46] Cha Zhang, D. Florencio, and Zhengyou Zhang. Why does phat work well in low noise, reverberative environments? In *Proceedings of ICASSP 2008*, pages 2565 –2568, 31 2008-april 4 2008.

Chapter 6

Conference Paper on GCC-PHAT model for Calibration in Diffusse Noise: *Novel GCC-PHAT Model in Diffuse Noise for Microphone Array Pairwise Distance based Calibration*

Publication reference:

- J. Velasco, M. J. Taghizadeh, A. Asaei, H. Bourlard, C. J. Martín-Arguedas, J. Macias-Guarasa, and D. Pizarro, "Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2669–2673

NOVEL GCC-PHAT MODEL IN DIFFUSE SOUND FIELD FOR MICROPHONE ARRAY PAIRWISE DISTANCE BASED CALIBRATION

Jose Velasco[1], Mohammad J. Taghizadeh[2,3,4], Afsaneh Asaei[2], Hervé Bourlard[2,3],
Carlos J. Martín-Arguedas[1], Javier Macias-Guarasa[1], Daniel Pizarro[1]
[1]Department of Electronics, University of Alcalá, Alcalá de Henares, Spain
[2]Idiap Research Institute, Martigny, Switzerland
[3]École Polytechnique Fédérale de Lausanne, Switzerland
[4]Huawei European Research Center, Munich, Germany
{jose.velasco, cj.martin, macias, pizarro}@depeca.uah.es, mohammad.taghizadeh@huawei.com,{afsaneh.asaei, herve.bourlard}@idiap.ch

ABSTRACT

We propose a novel formulation of the generalized cross correlation with phase transform (GCC-PHAT) for a pair of microphones in diffuse sound field. This formulation elucidates the links between the microphone distances and the GCC-PHAT output. Hence, it leads to a new model that enables estimation of the pairwise distances by optimizing over the distances best matching the GCC-PHAT observations. Furthermore, the relation of this model to the coherence function is elaborated along with the dependency on the signal bandwidth. The experiments conducted on real data recordings demonstrate the theories and support the effectiveness of the proposed method.

Index Terms— Generalized cross correlation, Phase transform, Diffuse sound field, Pairwise distance estimation, Microphone array calibration

1. INTRODUCTION

Microphone arrays are widely used to enable high-quality distant audio acquisition. They are an essential part of a plethora of distant technologies ranging from source localization and separation to distant speech recognition [1, 2, 3] and from sound field analysis and monitoring to virtual reality and surveillance [4, 5]. A fundamental pre-processing step to enable the array of microphones to function in synergy consists of the gain, clock and position calibration. In this paper we address the problem of microphone array position calibration or extracting the relative geometry or the shape of the microphone array.

The prior art often rely on activation of known signals to estimate the pairwise microphone distances. This approach is referred to as self-calibration. Sachar et al. [6] presented an experimental setup using a pulsed acoustic excitation generated by five domed tweeters. The transmit times between speakers and microphones were used to find the relative geometry. Raykar et al. [7] used a maximum length sequence or chirp signal in a distributed computing platform. The time difference of arrival of the microphone signals were then computed by cross-correlation and used for estimating the microphone locations. Since the original signal is known, these techniques are robust to noise and reverberation.

In an alternative approach to alleviate the requirement for a known signal, Chen et al. [8] introduced an energy-based method for joint microphone calibration and source localization. The energy of the signal is computed and a nonlinear optimization problem is formulated to perform maximum likelihood estimation of the source-sensor positions. This method requires several active sources for accurate localization and calibration. Pollefeys and Nistre proposed a method for direct joint source and microphone localization which requires matrix factorization and solving linear equations [9]. In a different approach, McCowan et al. [10] proposed a calibration method which does not require activation of a particular signal. This approach relies on the characteristics of a diffuse sound field. A diffuse field can be roughly described as an acoustic field where the signals propagate with equal probability in all directions with the same power. The diffuse field is verified for meeting rooms and car environments [11, 12] and it enables application of well-defined mathematical models for analysis of the acoustic field recordings. A particular property related to diffuse field recordings is the coherence function between pairwise microphone signals which is defined by a sinc function of the distance between the two microphones. Thereby, we can estimate the pairwise distances by least-squares fitting the computed coherence with the sinc function.

In this paper, we derive a new model based on generalized cross correlation with phase transform (GCC-PHAT) for a diffuse sound field. This model elucidates the links between the output of GCC-PHAT and the distance between the microphone pairs. The relation between GCC-PHAT and the coherence has been previously discussed in [13, 14] where PHAT filtering is used as an estimator of the coherence between two signals. The global coherence field described in [15], has a virtually identical formulation to the steered response power with phase transform [16], which can be expressed in terms of GCC-PHAT [17]. Both rely on using the classical beamforming techniques in order to build an acoustic power map of the room, which has been reported in [18] to coincide with the maximum likelihood estimation of the position of the source under low noise and high reverberation conditions. In [19], a novel GCC-PHAT model is established for a point source, being validated with both synthetic and real data. Based on the statistical analysis model of a diffuse sound field, we derive an extension of the GCC-PHAT model for a diffuse field. We present the procedure for estimating the pairwise distance from the GCC-PHAT function of the microphone recordings and elaborate its relation to the coherence-based approach [10].

The rest of the paper is organized as follows: The definition of GCC-PHAT and its model for the point sources is stated in Section 2, showing its behavior with respect to the source direction of arrival and the model extension for a diffuse sound field. In Section 3, the procedure for pairwise distance estimation is presented and contrasted with the alternative technique based on coherence fitting. The experimental evaluation on real data recordings is conducted in Sec-

tion 4, and the conclusions are drawn in Section 5.

2. GCC-PHAT IN DIFFUSE SOUND FIELD

In this section, we explain the new GCC-PHAT model for a point-source that establishes the links between the microphone array geometry and the GCC-PHAT output. We derive its extension for a diffuse sound field.

2.1. Generalized Cross-Correlation

The generalized cross-correlation (GCC) has been widely used for time-difference-of-arrival estimation and it is the basis for many acoustic source localization algorithms. The GCC of the signals recorded by two microphones is defined as:

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Psi_{ij}(\omega)\, X_i(\omega)\, X_j^*(\omega)\, \mathrm{e}^{\mathrm{j}\omega\tau}\, \mathrm{d}\omega, \qquad (1)$$

where $X_i(\omega)$ and $X_j(\omega)$ denote the signals recorded by microphones i and j in Fourier domain; ω is the angular frequency, $[.]^*$ stands for the conjugate transpose operation, and $\mathrm{j} = \sqrt{-1}$. The weighting function $\Psi_{ij}(\omega)$ is designed to optimize a given performance criteria. Many different functions have been proposed in the literature depending on the context, and among all of them, the phase transform (PHAT), defined as:

$$\Psi_{ij}(\omega) = \frac{1}{|X_i(\omega)\, X_j^*(\omega)|} = \frac{1}{|X_i(\omega)|\,|X_j(\omega)|}\,, \qquad (2)$$

has been found to perform very well for acoustic localization in reverberant environments, leading to the GCC-PHAT method [20] (also known as the crosspower-spectrum phase [15]). The PHAT can be seen as a filter which discards the amplitude and preserves the phase of the signal. The advantage of using it is that no assumptions are made about the signal or room conditions, which are typically unknown. This procedure has received considerable attention due to its simplicity and robustness in real world scenarios [18].

2.2. Analytic Model for a Point Source

The authors of [19] derive an analytical model for accurately predicting the behavior of the SRP-PHAT power maps for wideband signals, taking into account both the room geometry and the microphone array topology. They also show that the model is independent of the spectral content of the recorded signals, for both anechoic and reverberant conditions.

We consider a scenario where a single source is present and generates a baseband signal with bandwidth ω_0, thus $X_i(\omega) = 0$, $\forall\omega > \omega_0$. Assuming a free-space propagation model and discarding the distance dependent attenuation which is not relevant to our purposes, the signal at microphone j can be represented as a time-shifted version of $X_i(\omega)$, i.e. $X_j(\omega) = X_i(\omega)\mathrm{e}^{-\mathrm{j}\omega\tau_p}$ where τ_p is the time-difference of arrival between the two microphones.

From the model proposed in [19], and considering the anechoic propagation case, it is easy to show that when GCC-PHAT is applied to the signals captured by the microphone array, the resulting correlation can be approximated as a sinc function ($\mathrm{sinc}(x) = \frac{\sin(x)}{x}$), through

$$\begin{aligned} R_{\text{PHAT}}^{\text{point-source}}(\tau,\tau_p) &\approx \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} \mathrm{e}^{\mathrm{j}\omega(\tau-\tau_p)}\,\mathrm{d}\omega \\ &= \frac{\omega_0}{\pi}\,\mathrm{sinc}\left(\omega_0\,(\tau-\tau_p)\right)\,. \end{aligned} \qquad (3)$$

It may be noted that τ_p depends on the position of the source signal and it is limited by the distance d between two microphones such that $\tau_p \in \left[\frac{-d}{c}, \frac{d}{c}\right]$ with c being the speed of sound.

2.3. Extension to the Diffuse Sound Field

A diffuse field is defined as an acoustic field consisting of a superposition of an infinite number of sound waves traveling with random phases and amplitudes such that the energy density is equivalent at all points. More precisely, all points in the field radiate equal power and random phase sound waves, with the same probability for all directions, and the field is homogeneous and isotropic [21]. The analytic studies to model the diffuse sound field often rely on the statistical approach by considering an infinite number of free propagation plane waves, referred to as the plane wave model. In the plane wave model, a diffuse field is characterized as the superposition of a large set of plane waves impinging from all directions.

The spatial uniformity in a diffuse field can be expressed through integration of waves arriving from all directions [22, 23]. For two microphones, integrating over all directions is equivalent to integrating over all possible time-differences of arrival $\tau_p \in \left[\frac{-d}{c}, \frac{d}{c}\right]$ [22]. Therefore, the GCC-PHAT obtained in a diffuse field can be approximated by the GCC-PHAT model for a single source through the integration of uncorrelated sources arriving uniformly at all possible time-differences of arrival:

$$\begin{aligned} R_{\text{PHAT}}^{\text{diffuse}}(\tau,d) &\approx \int_{-\frac{d}{c}}^{\frac{d}{c}} R_{\text{PHAT}}^{\text{point-source}}(\tau,\tau_p)\;\frac{c}{2d}\,\mathrm{d}\tau_p \\ &= \frac{c}{2\pi d}\left(\mathrm{Si}\left(\omega_0\,(\tau+d/c)\right) - \mathrm{Si}\left(\omega_0\,(\tau-d/c)\right)\right), \end{aligned} \qquad (4)$$

where $\mathrm{Si}(x) = \int_0^x \mathrm{sinc}(t)\,\mathrm{d}t$ is the *sine integral*. The model expressed by (4) only depends on the distance between microphones d, and the signal bandwith ω_0. Furthermore, for large enough ω_0, the model can be approximated by a scaled version of the rectangular function:

$$\Pi\left(\frac{c\tau}{2d}\right) = \begin{cases} 0 & : |\tau| > \frac{d}{c} \\ \frac{1}{2} & : |\tau| = \frac{d}{c} \\ 1 & : |\tau| < \frac{d}{c} \end{cases} \qquad (5)$$

Fig. 1 demonstrates an example of the model and the real data measurements, for two different bandwidth values. Note that the values of $|\tau| > \frac{d}{c}$ do not provide relevant information about the distance between the two microphones while they nevertheless introduce some noise. Hence, it is easy to increase the signal-to-noise ratio by discarding those τs which do not have physical meaning based on the prior knowledge on the dimensions of the room or the physical setup.

3. MICROPHONE ARRAY CALIBRATION

In this section, we explain how the model of GCC-PHAT in diffuse sound field can be exploited to estimate the pairwise distance between two microphones for microphone array geometry calibration.

3.1. Distance Estimation Based on GCC-PHAT Model

The GCC-PHAT function for the signals of two microphones is obtained from (1)–(2) thus $R_{\text{PHAT}}(\tau)$ denotes the output based on the real data recordings. From the GCC-PHAT model expressed in (4), the distance between microphones can be estimated by fitting the model as:

$$\hat{d} = \underset{d,K}{\arg\min} \sum_{\tau=-\tau_{\max}}^{\tau_{\max}} \left(K\, R_{\text{PHAT}}(\tau) - R_{\text{PHAT}}^{\text{diffuse}}(\tau,d)\right)^2\,, \qquad (6)$$

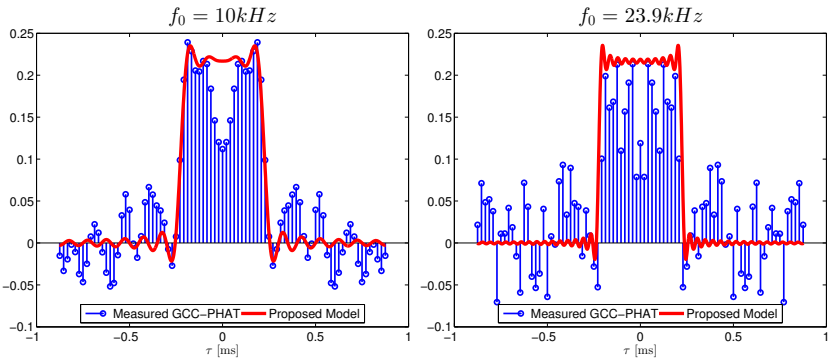


Fig. 1. The proposed GCC-PHAT model (4) contrasted with the measured GCC-PHAT on real data recordings in a diffuse sound field recorded at the room described in Section 4.1. The dependency on the signal bandwidth is demonstrated: the left graphic uses $f_0 = 10$ kHz and the right one uses $f_0 = 23.9$ kHz. We can see that for larger f_0 the model gets closer to the ideal case expressed in Eq. (8). Moreover we can see that the model fitting is better for smaller f_0 which is related to the fundamental limitations of a diffuse sound field for pairwise distance estimation [24].

where τ is discretized according to the sampling frequency and $\tau_{\max} = \frac{d_{\max}}{c}$. $d_{\max}$ indicates the expected maximum pairwise distance between any two microphones in the array, and can be estimated using geometrical considerations regarding the maximum room dimensions and the expected array geometry and locations. The additional parameter $K > 0$ is necessary since, in real scenarios, the model overestimates the amplitude of the correlation, which is lower due to the noise. Stacking the components for all values of τ, we obtain:

$$\mathbf{R}_{\text{PHAT}} \triangleq [R_{\text{PHAT}}(-\tau_{\max}), \dots, R_{\text{PHAT}}(\tau_{\max})]\,,$$
$$\mathbf{R}^{\text{diffuse}}_{\text{PHAT}}(d) \triangleq \left[R^{\text{diffuse}}_{\text{PHAT}}(-\tau_{\max}, d), \dots, R^{\text{diffuse}}_{\text{PHAT}}(\tau_{\max}, d)\right],$$

and after being Euclidean normalized, we obtain $\hat{\mathbf{R}}_{\text{PHAT}}$ and $\hat{\mathbf{R}}^{\text{diffuse}}_{\text{PHAT}}$. It it straightforward to show that, for discrete τ, minimizing the quadratic error $\left(K\, R_{\text{PHAT}}(\tau) - R^{\text{diffuse}}_{\text{PHAT}}(\tau, d)\right)^2$ is equivalent to minimizing the angle between the normalized vectors. Hence, denoting the inner product between two unit vectors by $\langle ., . \rangle$, we can rewrite Eq. (6) as:

$$\hat{d} = \arg\max_{d} \left\langle \hat{\mathbf{R}}_{\text{PHAT}}, \hat{\mathbf{R}}^{\text{diffuse}}_{\text{PHAT}}(d) \right\rangle \tag{7}$$

Given all the (offline-calculated) unitary vectors $\hat{\mathbf{R}}^{\text{diffuse}}_{\text{PHAT}}(d)$, the one that is better aligned with the $\hat{\mathbf{R}}_{\text{PHAT}}$ computed from the data can be found efficiently, indicating an estimate of the pairwise distance d.

3.2. Relation to the Coherence

The GCC-PHAT and coherence are two terms which are closely interconnected [13, 14]. The coherence of two signals is defined as the cross spectrum normalized by the square roots of the auto spectra. It has been shown that the real-part of the coherence of the signals at each frequency in a diffuse sound field is a $\text{sinc}\left(\frac{\omega d}{c}\right)$ function of the microphone distances [25]. This property is exploited by McCowan et al. to estimate the microphone pairwise distances [26].

In this section we show that the model introduced in equation (4) is, in fact, a low-pass filtered version of the inverse Fourier transform of the coherence-based approach [26]. Based on Eqs. (3) and (4), the GCC-PHAT model for the diffuse sound field can be written as:

$$\begin{aligned} R^{\text{diffuse}}_{\text{PHAT}}(\tau, d) &\approx \int_{-\omega_0}^{\omega_0} \int_{-\frac{d}{c}}^{\frac{d}{c}} \mathrm{e}^{\mathrm{j}\omega(\tau-\tau_p)}\, \frac{c}{2d}\, \mathrm{d}\tau_p\, \mathrm{d}\omega \\ &= \int_{-\omega_0}^{\omega_0} \frac{c}{2d\omega \mathrm{j}} \left(\mathrm{e}^{\mathrm{j}\omega(t+\frac{d}{c})} - \mathrm{e}^{\mathrm{j}\omega(t-\frac{d}{c})}\right) \mathrm{d}\omega \\ &= \int_{-\omega_0}^{\omega_0} \text{sinc}\left(\frac{\omega d}{c}\right)\, \mathrm{e}^{\mathrm{j}\omega}\, \mathrm{d}\omega \end{aligned} \tag{8}$$

Hence, we can see that the GCC-PHAT model for a diffuse sound field is the Fourier transform of the sinc (real-part of the coherence) ideally filtered at ω_0. Since the proposed model is the inverse Fourier transform of the coherence-based model, removing high values of τ in the GCC-PHAT calculation, implies removing fast changes in the coherence and lead to denoising the coherence; Fig. 2 demonstrates an example of the denoising effect achieved via suppressing the time coefficients corresponding to $\tau > \tau_{\max}$. As we will see during the experimental evaluation presented in Section 4, the GCC-PHAT model in a diffuse sound field outperforms the coherence-based approach [26], while improving the computational cost.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of the proposed technique for pairwise distance estimation using real data recordings collected at the Idiap smart meeting room.

4.1. Acoustic Recording Setup

We use the geometrical setup of the MONC corpus to record the sound field in a meeting room [27]. The enclosure is a $8 \times 5.5 \times 3.5\,\text{m}^3$ rectangular room and it is moderately reverberant. It contains a centrally located $4.8 \times 1.2\,\text{m}^2$ rectangular table. Nine microphones are located on a planar area parallel to the floor at a height of 1.15 m: Eight of them are located on a circle with diameter 20cm and one microphone is at the origin. The microphones are Sennheiser MKE-2-5-C omnidirectional miniature lapel microphones. The floor of the room is covered with carpet and surrounded with plaster walls and two big windows; the room is mildly reverberant with a reverberation time less than 200 ms. The room is almost silent and no

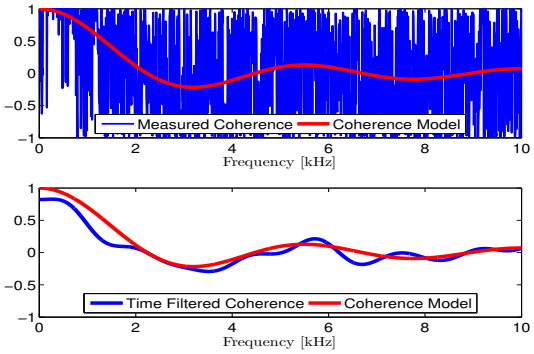


Fig. 2. The frame-based coherence measured using the real data and the theoretical sinc model in the original form (top) and after time filtering (bottom) based on suppression of the GCC-PHAT output at the large time intervals that do not correspond to the physical setup.

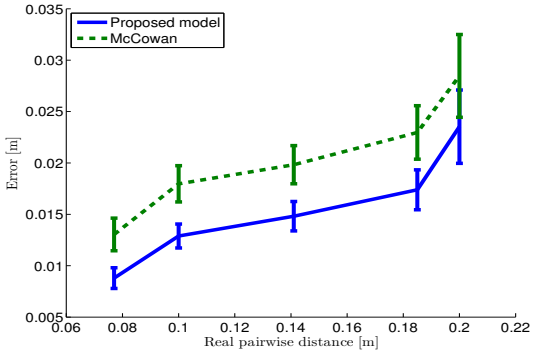


Fig. 3. Comparison between the average error in distance estimation using the proposed GCC-PHAT model (7) and the McCowan's coherence-based method [10].

sound source is generated; there is ambient noise due to the street and computer fans. The sampling rate is 48 kHz. The experiments are conducted using $c = 343$ m/s that corresponds to 20° Celsius temperature of the room.

4.2. Analysis Parameters

The recordings are processed frame by frame in frames of 4096 samples (85.3 ms) after applying the Tukey window. The FFT is calculated using 8192 samples (after zero-padding). The maximum distance between microphones was restricted to $1.5m$, so that all τ in GCC-PHAT corresponding to longer distances were not considered. The set of possible distances are discretized within the range of $[0.05, 1.5]$ m with one millimeter resolution.

Since the diffuse noise is expected to be broadband and with equal power in all frequencies, $\omega_0 = 2\pi f_0$ has been in fact determined by the antialiasing filter ($f_0 = 23.9$ KHz). A more restrictive filtering allows a better fitting, as demonstrated in Fig. 1.

4.3. Pairwise Distance Estimation Performance

Fig. 3 shows the estimation error of pairwise distances for the two models; the bars represent the 99% confidence interval, assuming a normal distribution. The improvement of the proposed model in terms of pairwise distance estimation is statically significant, but it does not lead to better results in the calibration of the position of the microphones based on multidimensional scaling method [28].

4.4. Numerical Approximation for the Proposed Model

The mathematical approximation is suitable as Matlab® provides a symbolic implementation for the sine integral which can be sometimes quite slow. We suggest using the numerical approximation described in [29, p. 231]:

$$\mathrm{Si}\,(x) \approx \begin{cases} \sum_{n=0}^{N-1} \frac{(-1)^n x^{2n+1}}{(2n+1)(2n+1)!} & : |x| \leq 1 \\ \frac{\pi}{2} - f(x)\cos(x) - g(x)\sin(x) & : x > 1 \\ f(-x)\cos(x) - g(-x)\sin(x) - \frac{\pi}{2} & : x < 1 \end{cases} \tag{9}$$

which has a low error ($|\epsilon(x)| < \max\{\frac{1}{(2N+1)2!}, 3 \times 10^{-7}\}$), and it can speed up the implementation of the proposed model. Functions $f(x)$ and $g(x)$ are calculated as [1] :

$$f(x) = \frac{1}{x}\left(\frac{x^8 + a_1x^6 + a_2^4 + a_3x^2 + a_4}{x^8 + b_1x^6 + b_2^4 + b_3x^2 + b_4}\right) \tag{10a}$$

$$g(x) = \frac{1}{x^2}\left(\frac{x^8 + c_1x^6 + c_2^4 + c_3x^2 + c_4}{x^8 + d_1x^6 + d_2^4 + d_3x^2 + d_4}\right) \tag{10b}$$

The above approximation speeds up the process more than one million times. The time that it takes to perform pairwise distance estimation using each frame is 40 times faster than real time. The new GCC-PHAT model is also 30 times faster than the alternative coherence-base approach.

5. CONCLUSIONS

In this paper, a new model for GCC-PHAT in diffuse sound field is proposed which establishes the links between GCC-PHAT output and the microphone array geometry. To estimate the pairwise distances, the GCC-PHAT is computed for a pair of microphone signals and the distance that generates the best fitting model is estimated. It was shown that this model is in fact equivalent to an inverse Fourier transform of an ideally filtered coherence of the two signals. The experiments conducted on real data recordings demonstrate the effectiveness of the proposed approach for pairwise distance estimation. Furthermore, it suggests a simple denoising scheme for the coherence function via suppression of the GCC-PHAT activation at the time intervals which do not meet the physical constraints. The model was shown to perform significantly faster than the coherence-based counterpart and it is applicable for real time calibration setups.

6. ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R), and by the FPU Grants Program of the University of Alcalá. Afsaneh Asaei acknowledges the SNSF 200021-153507 grant on PHASER project.

[1] $a_1 = 38.027264$, $a_2 = 265.187033$, $a_3 = 335.677320$, $a_4 = 38.102495$, $b_1 = 40.021433$, $b_2 = 322.624911$, $b_3 = 570.236280$, $b_4 = 157.105423$, $c_1 = 42.242855$, $c_2 = 302.757865$, $c_3 = 352.018498$, $c_4 = 21.821899$, $d_1 = 48.196927$, $d_2 = 482.485984$, $d_3 = 1114.978885$ and $d_4 = 449.690326$.

7. REFERENCES

[1] H. T. Do, *Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array*, Ph.D. thesis, Brown University, 2011.

[2] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 3, pp. 620–633, 2014.

[3] Afsaneh Asaei, *Model-based Sparse Component Analysis for Multiparty Distant Speech Recognition*, Ph.D. thesis, École Polytechnique Fédéral de Lausanne (EPFL), 2013.

[4] M. Mattila V. Veijanen V. Pulkki T. Hiekkanen, T. Lempiainen, "Reproduction of virtual reality with multichannel microphone techniques," in *Proceeding of 122nd AES Convention*, 2007.

[5] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, E Antonacci, and Augusto Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, 2007, pp. 21–26.

[6] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13(1), 2005.

[7] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13(1), 2005.

[8] M. Chen, Z. Liu, L. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad-hoc microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[9] Marc Pollefeys and David Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data.," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 2445–2448.

[10] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on Audio,Speech and Language Processing*, vol. 16(3), 2008.

[11] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *IEEE workshop on Hands-free Speech Communication and Microphone Arrays*, 2011.

[12] J. Bitzer, K. U. Simmer, and K. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[13] Maurizio Omologo and Piergiorgio Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, vol. 2, pp. II–273.

[14] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer, "Speaker localization based on oriented global coherence field," in *Proceedings of Interspeech*, 2006, vol. 7, p. 8.

[15] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 288–292, 1993.

[16] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone Arrays*, chapter Robust Localization in Reverberant Rooms, pp. 157–180, 2001.

[17] J.H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[18] Cha Zhang, D. Florencio, and Zhengyou Zhang, "Why does phat work well in low noise, reverberative environments?," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 2565 –2568.

[19] Jose Velasco, Carlos J. Martín-Arguedas, Javier Macias-Guarasa, Daniel Pizarro, and Manuel Mazo, "Proposal and validation of an analytical generative model of srp-phat power maps in reverberant scenarios," Tech. Rep. GEINTRA-RR-1-2014, GEINTRA Research Group, Department of Electronics, University of Alcala, Spain, November 2004, http://www.geintra-uah.org/RR-14-01.pdf.

[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320 – 327, aug 1976.

[21] T.J. Schultz, "Diffusion in reverberation rooms," *Journal of Sound and Vibration*, vol. 16(1), 1971.

[22] Boaz Rafaely, "Spatial-temporal correlation of a diffuse sound field," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3254–3258, 2000.

[23] Allan D Pierce et al., *Acoustics: an introduction to its physical principles and applications*, McGraw-Hill New York, 1981.

[24] Mohammad J. Taghizadeh, Philip N. Garner, and Hervé Bourlard, "Enhanced diffuse field model for ad hoc microphone array calibration," *Signal Processing*, vol. 101, 2014.

[25] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlations coefficients in reverberan sound fields," *Journal of the Acoustical Society of America*, vol. 27, 1955.

[26] Iain McCowan, Mike Lincoln, and Ivan Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 3, pp. 666–670, 2008.

[27] "The multichannel overlapping numbers corpus (MONC)," Idiap resources available online:, `http://www.cslu.ogi.edu/corpora/monc.pdf`.

[28] T. F. Cox and M. A. A. Cox, "Multidimensional scaling," *Chapman-Hall, 2001*.

[29] Milton Abramowitz, Irene A Stegun, et al., *Handbook of mathematical functions*, vol. 1, Dover New York, 1972.

Chapter 7

Journal Paper on TDOA Matrices: *TDOA Matrices: Algebraic Properties and their Application to Robust Denoising with Missing Data*

Publication reference:

- J. Velasco, D. Pizarro, J. Macias-Guarasa, and A. Asaei, "TDOA matrices: Algebraic properties and their application to robust denoising with missing data," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5242–5254, Oct 2016

5242 IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 64, NO. 20, OCTOBER 15, 2016

TDOA Matrices: Algebraic Properties and their Application to Robust Denoising with Missing Data

Jose Velasco, *Student Member, IEEE*, Daniel Pizarro, Javier Macias-Guarasa, *Member, IEEE*, and Afsaneh Asaei, *Senior Member, IEEE*

***Abstract*—Measuring the Time delay of Arrival (TDOA) between a set of sensors is the basic setup for many applications, such as localization or signal beamforming. This paper presents the set of TDOA matrices, which are built from noise-free TDOA measurements, not requiring knowledge of the sensor array geometry. We prove that TDOA matrices are rank-two and have a special SVD decomposition that leads to a compact linear parametric representation. Properties of TDOA matrices are applied in this paper to perform denoising, by finding the TDOA matrix closest to the matrix composed with noisy measurements. The paper shows that this problem admits a closed-form solution for TDOA measurements contaminated with Gaussian noise which extends to the case of having missing data. The paper also proposes a novel robust denoising method resistant to outliers, missing data and inspired in recent advances in robust low-rank estimation. Experiments in synthetic and real datasets show significant improvements of the proposed denoising algorithms in TDOA-based localization, both in terms of TDOA accuracy estimation and localization error.**

***Index Terms*—TDOA estimation, TDOA denoising, skew-symmetric matrices, matrix completion, missing data**

I. Introduction

TIME delay of arrival (TDOA) estimation is an essential pre-processing step for multiple applications in the context of sensor array processing, such as multi-channel source localization [1], self-calibration [2] and beamforming [3]. In all cases, performance is directly related to the accuracy of the estimated TDOAs [4].

Estimating TDOA in noisy environments has been subject of study during the last two decades [5]–[7], and is still an active area of research, benefiting from current advances in signal processing and optimization strategies [8]–[11].

Typically, the TDOA between a single pair of sensors is obtained by measuring the peak of the generalized cross-correlation (GCC) of the received signals on each sensor [12], which are assumed to be generated from a single source. Many factors, such as the spectral content of the signal, multipath propagation, and noise contribute to errors in the estimation of the TDOA.

J. Velasco, D. Pizarro, and J. Macias-Guarasa are with the Department of Electronics, Escuela Politécnica Superior, University of Alcalá, 28805 Alcalá de Henares, Spain.

A. Asaei is with Idiap Research Institute, Martigny, Switzerland.

E-mails: {jose.velasco, pizarro, macias}@depeca.uah.es, afsaneh.asaei@idiap.ch).

Manuscript received January 15, 2016; revised May 18, 2016 and June 30, 2016; accepted July 08, 2016. Date of publication July 21, 2016;

Given a set of sensors, TDOA measurements can be obtained for every possible pair of sensors. This is commonly known as the *full TDOA set* or *spherical set* [13]. This paper studies how to reduce noise and errors from the full TDOA set. The intuition behind this denoising is to exploit redundancy of the full TDOA set. For n sensors, the full set of $n(n-1)/2$ measurements can be represented by $n-1$ values, which are referred to as the *non-redundant set*. This problem has been studied in the past, showing that one can optimally obtain the *non-redundant set* when TDOA measurements are contaminated with additive Gaussian noise. This is known as the Gauss-Markov estimator [14]. However, in more realistic scenarios errors are not Gaussian and some of the TDOA measurements may contain outliers. In these cases the Gauss-Markov estimator performs poorly.

This paper presents the TDOA matrix, which is created by the arrangement of the full TDOA set inside a skew-symmetric matrix, and studies the algebraic properties of this matrix, showing that it has rank 2 and a SVD decomposition with $n-1$ degrees of freedom. Such matrices have been previously defined in the literature [15], but their properties and applications have not been studied until now.

These algebraic properties are used in this paper to perform denoising under different scenarios, which include the presence of missing TDOA measurements and outliers. These denoising algorithms are tested in the context of speaker localization with microphone arrays, using synthetic and publicly available real datasets. Our denoising algorithms are able to recover accurate TDOA values for high rates of missing data and outliers, significantly outperforming the Gauss-Markov estimator in those cases. All the proposed methods don't require knowledge of the sensor positions, so that they can also be used for calibration [2].

The main contributions of this work are threefold: *i)* Definition of the algebraic properties of TDOA matrices. *ii)* A closed-form solution for TDOA denoising for Gaussian noise and the presence of missing data. *iii)* Novel robust-denoising methods for handling additive correlated noise, outliers and missing data.

A. Notation

Real scalar values are represented by lowercase letters (e.g. δ). Vectors are by default arranged column-wise and are represented by lowercase bold letters (e.g. $\mathbf{x}$). Matrices are represented by uppercase bold letters (e.g. $\mathbf{M}$). Lower-case letters are reserved to define vector and set sizes (e.g. vector

$\mathbf{x} = (x_1, \cdots, x_n)^\top$ is of size n), and $\mathbf{x}^\top$ denotes transpose of vector $\mathbf{x}$. Calligraphic fonts are reserved to represent generic sets (e.g. $\mathcal{G}$) or functions applied to matrices (e.g. $\mathcal{P}(\mathbf{X})$). The l_2 norm $\|\cdot\|_2$ will be written by default as $\|\cdot\|$ for simplicity, and $\|\cdot\|_F$ is the Frobenius norm, while $|\cdot|$ is reserved to represent absolute values of scalars. The l_0 norm of a matrix, written $\|\cdot\|_0$, is defined as the number of non-zero elements of the matrix. $\mathbf{A} \circ \mathbf{B}$ is the Hadamard product between $\mathbf{A}$ and $\mathbf{B}$, defined as the entrywise multiplication of the corresponding matrices. $\mathrm{tr}(\cdot)$ is the trace function.

We also define the normalized unitary vector $\hat{\mathbf{1}}$ as $\hat{\mathbf{1}} = (1, \ldots, 1)^\top / \sqrt{n}$, and the null vector $\hat{\mathbf{0}}$ as $\hat{\mathbf{0}} = (0, \ldots, 0)^\top$, both of them having size n. Finally, $\mathbb{1} = n\, \hat{\mathbf{1}}\, \hat{\mathbf{1}}^\top$ is a $n \times n$ matrix with all elements equal to 1, $\mathbf{D}_\mathbf{x}$ is a $n \times n$ diagonal matrix where its main diagonal is the vector $\mathbf{x}$, and $\mathbf{I}$ is the identity matrix.

B. Paper Structure

The rest of the paper is distributed as follows. Section II describes the related work and Section III the problem statement. In section IV TDOA matrices are described along with a derivation of their properties. TDOA denoising in Gaussian noise case is addressed in section V, also providing a closed-form solution. In sections VI and VII we propose novel algorithms for robust handling of noise and missing data, respectively. Section VIII combines the proposals of the previous two sections into an unified algorithm. We also provide an extensive experimentation to validate the proposed algorithms using both synthetic (Section IX) and real data (Section X). Finally, conclusions are drawn in section XI.

II. Related Work

TDOA estimation is an essential first step for multiple applications related to *1)* localization, *2)* self-calibration and *3)* beamforming (among others):

1) Localization: widely used in radar, sonar and acoustics, since no synchronization between the source and sensor is needed. The TDOA information is combined with knowledge of the sensors' positions to generate a Maximum Likelihood spatial estimator made from hyperboloids intersected in some optimal sense. A linear closed-form solution of the former problem, valid when the TDOA estimation errors are small, is given in [16].

2) Self-calibration: since knowing the position of sensors is mandatory for localization techniques, some strategies have been also proposed in order to calibrate them using only TDOA measurements. In [2], [17], the TDOA problem is converted in a Time of Arrival (TOA) problem estimating the departure time of signals. Then, self-calibration techniques for TOA can be employed. The main drawback of this approach is that the conversion step from TDOA to TOA is very sensitive to outliers and correlated noise.

3) Beamforming: precise TDOA estimations is also critical for beamforming techniques and its applications. In [3], for example, additional steps are proposed for selecting the appropriate TDOA value among the correlation peaks, and also dealing with TDOA outliers. These steps include a Viterbi decoding based algorithm which maximizes the continuity of the TDOA estimations in several frames. However, the TDOA selection criteria is just based on their distance to surrounding TDOA values and their GCC-PHAT values, thus not attempting to benefit from the actual redundancy of the TDOA measurements.

Hence, an accurate estimation of TDOA is essential for a good performance of any of the former applications based on these measurements.

Typically, when only two sensors are employed, the peak of the generalized cross-correlation (GCC) function of the signals of two sensors is a good estimator for the TDOA, for reasonable noise and reverberation levels [12].

When more than two sensors are used ($n > 2$), there are $n(n-1)/2$ different TDOA measurements from all possible pairs of sensors, forming the *full TDOA set* or *spherical set* [13]. However, all those TDOA measurements are redundant. In fact, usually one sensor is considered the reference sensor, and only the subset of $n-1$ TDOA measurements which involve that sensor are considered. That *non-redundant set* is the set of measurement used by the majority of TDOA-based positioning algorithms proposed in the literature [16], [18]–[22]. Nevertheless, an optimal (denoised) version of the *non-redundant set* can be estimated from the redundant set using a Bayesian Linear Unbiased Estimator (BLUE), also known as the Gauss-Markov estimator [14].

A closed-form solution for the BLUE estimator is provided in [23], also proving that it is equal to the standard least squares estimator, and that it reaches the Cramer-Rao lower bound for positioning estimation. However, all the results in that work are based on the assumption of additive Gaussian noise, which is unrealistic in many practical applications [24], and doesn't yield good results when correlated noise is present as a consequence, for instance, of multipath propagation. Additionally, the experimental results shown in their work are only applied to synthetic data, thus not allowing to assess the performance of their proposal in real scenarios (in section X we show the limitations of their method when evaluated on real data).

A least-squares solution to TDOA denoising is proposed in [25]. It is based on projecting the *non-redundant set* of TDOA measurements into a set of "feasible" bivectors (rank 2, antisymmetric tensors) that show the same geometric properties of TDOA matrices. This denoising is also optimal for Gaussian noise but as in [24] the experimental analysis is based on simulated data and it does not cope with missing data or the presence of outliers in the TDOA measurements.

Periodicity in correlated signals, coherent noise and multi-path due to reverberation are the major sources of non-Gaussian error in TDOA estimation. Different approaches have been proposed to deal with them. A basic method consists in making the GCC function more robust, de-emphasizing the frequency-dependent weighting. The Phase Transform (PHAT) [26] is one example of this procedure which has received considerable attention as the basis of acoustic source localization systems due to its robustness in real world scenarios [27], [28]. Other approaches are based in blind estimation of multi-path (room impulse response) [29] but they need a

good initialization to perform well.

Some previous works have also proposed more complicated structures in order to represent TDOA redundancy, while not imposing strong assumptions on the noise distribution. In [30] a representation based on graphs allows to disambiguate if a peak in correlation was generated by the direct path or by reverberation applying an efficient search algorithm among all possible combinations. However, they do not explicitly attempt to provide improved TDOA estimations by exploiting their redundancy.

Also different matrix representations have been used in the bibliography regarding TDOA formulation. For example [31] uses a representation slightly different to the TDOA matrices we describe here, but such representation does not have the algebraic properties that TDOA matrices have, and their authors do not address an study in this sense.

So, to the best of our knowledge, there are no previous reported work dealing with improving TDOA estimations by exploiting their redundancy, while not imposing Gaussian noise restrictions, not requiring the sensor positions, and being able to deal with the presence of outliers and missing measurements (errors that will severely impact the performance of applications based on TDOA measurements). In this paper we show that TDOA matrices are a powerful tool that combined with recent advances in robust low-rank estimation, are able to generate novel solutions for these problems.

III. Problem Statement

Hereafter, we assume only one source located at the position $\mathbf{r} = (r_x, r_y, r_z)^\top$, and n sensors synchronized between them and placed in different positions $\mathbf{s}_i = (s_{ix}, s_{iy}, s_{iz})^\top$, $i \in [1, n]$.

Given this setup, we will assume that the source is emitting an unknown signal $x(t)$. Then, the signal received by the sensor i, $x_i(t)$, is without loss of generality, a delayed and attenuated version of $x(t)$ (direct propagation) in addition to a signal $g_i(t)$ which summarizes all the adverse effects, i.e. noise, interference, multipath, etc. Thus, $x_i(t) = x(t - \tau_i) + g_i(t)$, where $\tau_i = \|\mathbf{r} - \mathbf{s}_i\|_2 / c$ is the time of arrival (TOA) of the signal $x(t)$ at the sensor $\mathbf{s}_i$, being c the propagation speed.

Assuming that TOA cannot be estimated directly, the time delay of arrival (TDOA) between the sensors i and j is estimated by correlating the received signals $x_i(t)$ and $x_j(t)$ (typically using the Generalized Cross-Correlation GCC [26]).

IV. TDOA matrices

In this section we define TDOA matrices, and develop their main properties. In a nutshell, given any TDOA matrix $\mathbf{M}$, we show that: *i)* $\mathbf{M}$ is rank 2 (Theorem 1), *ii)* $\mathbf{M}$ can be decomposed as $\mathbf{M} = \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right)$ with $\mathbf{x} = \mathbf{M}\,\hat{\mathbf{1}}$ (Lemma 1) and *iii)* the previous decomposition is bijective (Theorem 2).

These properties are the foundations of the denoising algorithms that we present in sections V and VI, and the missing data recovery proposal described in section VII, plus their combination described in section VIII.

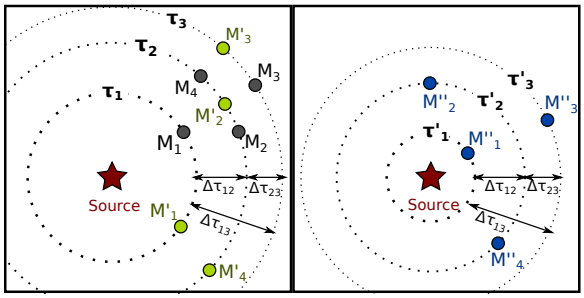


Fig. 1: Example of three different geometrical configurations (grey, green and blue) of 4 sensor with identical TDOA matrix.

A. Definition of TDOA matrices

Definition 1. *A TDOA matrix* $\mathbf{M}$*, is a* $(n \times n)$ *skew-symmetric matrix where the element* (i, j) *is the time difference of arrival (TDOA) between the signals arriving at sensor* i *and sensor* j*:*

$$\mathbf{M} = \{\Delta\tau_{ij}\} = \begin{pmatrix} 0 & \Delta\tau_{12} & \cdots & \Delta\tau_{1n} \\ \Delta\tau_{21} & 0 & \cdots & \Delta\tau_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta\tau_{n1} & \Delta\tau_{n2} & \cdots & 0 \end{pmatrix} \tag{1}$$

with $\Delta\tau_{ij} = (\tau_i - \tau_j)$*, where* τ_i *is the time of arrival of the signal* $x(t)$ *at the sensor* $\mathbf{s}_i$*.*

We will also express $\mathbf{M}$ in terms of its columns as $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_n)$, being $\mathbf{m}_i = (\Delta\tau_{1i}, \Delta\tau_{2i}, \ldots, \Delta\tau_{ni})^\top$.

We denote as $\mathcal{M}_T(n)$ to the set of TDOA matrices of size $n \times n$.

Notice that there is a bijection between the full TDOA set and the corresponding TDOA matrix. Nevertheless expressing TDOA measurements as a matrix has important advantages, that we will discover throughout this article.

Note also that in the former definition, knowing the sensor array geometry is not required. For a given geometry, all the feasible TDOA matrices (those that are consistent with that particular geometry) are a subset of $\mathcal{M}_T(n)$. Studying the properties of such subset is out of the scope of this paper and the interested reader can refer to [8], [9], [32] for further details.

Additionally, given a particular TDOA matrix, there are infinite number of sensors geometries which match with it. Left side of Fig. 1 shows that, given a set of TOAs (τ_1,..., τ_n) compatible with the set of TDOA measurements, the microphones can be situated in any place along the circumference (sphere in the 3D case) with center in the source (dotted lines), preserving its correspondent TOA (and therefore, its TDOA). Right side of Fig. 1 shows that there are an infinite number of TOA sets that comply with a given set of TDOA measurements.

B. Rank of TDOA matrices

Theorem 1. *Let* $\mathbf{M} \in \mathcal{M}_T(n)$*, then* $\mathbf{M}$ *is rank 2 or 0 (trivial case).*

Proof: The matrix $\mathbf{M}$ can be expressed as:

$$\mathbf{M} = \mathbf{T} - \mathbf{T}^\top, \tag{2}$$

where $\mathbf{T}$ is a rank 1 matrix defined as:

$$\mathbf{T} = \begin{pmatrix} \tau_1 & \cdots & \tau_1 \\ \vdots & \ddots & \vdots \\ \tau_n & \cdots & \tau_n \end{pmatrix}, \tag{3}$$

Applying the well known inequality:

$$\operatorname{rank}(\mathbf{A}+\mathbf{B}) \leq \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}), \tag{4}$$

we can deduce that $\operatorname{rank}(\mathbf{M}) \leq 2$.

Moreover, since the rank of any skew-symmetric matrix must be even, rank 1 is not feasible. So we can conclude that, excepting the case that $\mathbf{M}$ is the zero matrix (trivial case), the rank of $\mathbf{M}$ is 2. This completes the proof. ∎

Note that the formerly referred as 'trivial case' only occurs when the time of arrival is the same for all the sensors, i.e sensors are geometrically placed on a sphere with center in the source.

Rank deficiency of TDOA matrices means that their rows and columns are linearly dependent. That is consistent with the fact that, in the noise-free case, the full TDOA set can be generated from the *non-redundant set* using linear equations [23]. In fact, in a TDOA matrix, the column j is the TDOA *non-redundant set* when the sensor j is the reference for TDOA measurements. Hereafter, and without loss of generality, we will consider the first sensor as the reference for the *non-redundant set*.

C. Bijective mapping for TDOA matrices

Next, we show a particular representation of TDOA matrices with $n-1$ parameters that describes their spectral properties and also forms the base for our denoising algorithms.

Theorem 2. *Let $H \subset \mathbb{R}^n$ be a $n-1$ dimensional subspace of $\mathbb{R}^n$ such that $\operatorname{span}(\hat{1}) \nsubseteq H$, then there always exits an isomorphism between $\mathcal{M}_T(n)$ and H of form:*

$$\begin{aligned} \phi_H : H &\longrightarrow \quad \mathcal{M}_T(n) \\ \mathbf{z} &\longmapsto \mathbf{z}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{z}^\top \end{aligned}$$

Proof: Theorem 2 states that a bijective linear map exists between $\mathcal{M}_T(n)$ an a $n-1$ dimensional subset of $\mathbb{R}^n$. Since the matrix $\mathbf{T}$ in (3) can be rewritten as $\mathbf{T} = \mathbf{z}\,\hat{\mathbf{1}}^\top$, where $z = (\tau_1, \cdots, \tau_n)^\top$, we can define the following linear map:

$$\begin{aligned} \phi : \mathbb{R}^n &\longrightarrow \quad \mathcal{M}_T(n) \\ \mathbf{z} &\longmapsto \mathbf{z}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{z}^\top \end{aligned}$$

that is clearly surjective but not injective. That is, any vector $\mathbf{z}' = \mathbf{z} + \alpha\,\hat{\mathbf{1}}$ represents the same TDOA matrix. Indeed, the kernel of ϕ is the linear subspace of $\mathbb{R}^n$ generated by $\hat{\mathbf{1}}$.

Since we are looking for an isomorphism, we will restrict the domain of ϕ to ensure injectivity. It yields ϕ_H, where $H \subset \mathbb{R}^n$ is a hyperplane of $\mathbb{R}^n$ not containing $\ker(\phi)$. Note that ϕ_H is bijective as $H \subset \mathbb{R}^n$ keeps the surjectivity of ϕ and, since $\ker(\phi_H) = \{\mathbf{0}\}$, the function is also injective. ∎

Hereafter, we will only consider the particular case of ϕ_H for the hyperplane $H = \ker(\phi)^\perp$. It yields the following expression for any $M \in \mathcal{M}_T(n)$:

$$\mathbf{M} = \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right) \quad , \quad \mathbf{x} \perp \hat{\mathbf{1}} \quad , \mathbf{x} \in \mathbb{R}^n \tag{5}$$

Choosing $\mathbf{x}$ perpendicular to $\hat{\mathbf{1}}$ is very convenient, since it simplifies ϕ_H^{-1} as shown in the Corollary 2.1, and it can also be used for calculate the singular value decomposition (SVD) of any $M \in \mathcal{M}_T(n)$, as discussed in IV-C1. This leads to a parametric representation of $\mathbf{M}$ that has important properties that we will exploit later for TDOA denoising.

Corollary 2.1. *Given any $\mathbf{M} \in \mathcal{M}_T(n)$, the corresponding vector $\mathbf{x} \in \mathbb{R}^n$, perpendicular to $\mathbf{1}$, can be calculated as:*

$$\mathbf{x} = \mathbf{M}\,\hat{\mathbf{1}}. \tag{6}$$

Corollary 2.2. *$\mathbf{M} \in \mathcal{M}_T(n)$ can also be expressed as:*

$$\mathbf{M} = \frac{1}{\sqrt{n}}\left(\mathbf{D}_\mathbf{x}\,\mathbb{1} - \mathbb{1}\,\mathbf{D}_\mathbf{x}\right) \quad , \hat{\mathbf{1}} \perp \hat{\mathbf{x}}, \tag{7}$$

since $\mathbf{x}\,\hat{\mathbf{1}}^\top = \mathbf{D}_\mathbf{x}\,\mathbb{1}/\sqrt{n}$ and $\hat{\mathbf{1}}\,\mathbf{x}^\top = \mathbb{1}\,\mathbf{D}_\mathbf{x}/\sqrt{n}$,

1) Singular Value Decomposition: Because $\mathbf{M} \in \mathcal{M}_T(n)$ is a skew-symmetric matrix of rank 2, it has the following singular value decomposition (SVD) [33, Supplementary material]:

$$\mathbf{M} = (\hat{\mathbf{u}}_2, -\hat{\mathbf{u}}_1)\begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)^\top = \sigma\,(\hat{\mathbf{u}}_2, -\hat{\mathbf{u}}_1)\,(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)^\top, \tag{8}$$

where $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ are orthonormal vectors and $\sigma \geq 0$. Note that the SVD decomposition of $\mathbf{M}$ is not unique. Given any orthogonal 2×2 matrix $\mathbf{R}$, the vectors $(\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2) = (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)\,\mathbf{R}$ also represent a valid SVD decomposition:

$$\mathbf{M} = \sigma\,(\hat{\mathbf{v}}_2, -\hat{\mathbf{v}}_1)\,(\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2)^\top = \sigma\,(\hat{\mathbf{u}}_2, -\hat{\mathbf{u}}_1)\,(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)^\top. \tag{9}$$

Among all possible SVD decompositions of $\mathbf{M}$ note that Theorem 2 guarantees that always exists one where $\hat{\mathbf{u}}_1 = \hat{\mathbf{1}}$. In such case SVD decomposition can be computed from (5) as stated in the following theorem:

Lemma 1. *Given $\mathbf{M} \in \mathcal{M}_T(n)$, it admits the following SVD decomposition:*

$$\mathbf{M} = \sigma\left(\hat{\mathbf{u}}, -\hat{\mathbf{1}}\right)\left(\hat{\mathbf{1}}, \hat{\mathbf{u}}\right)^\top \quad \textit{with} \quad \hat{\mathbf{u}} = \frac{\mathbf{M}\,\hat{\mathbf{1}}}{\|\mathbf{M}\,\hat{\mathbf{1}}\|} \quad \sigma = \|\mathbf{M}\,\hat{\mathbf{1}}\| \tag{10}$$

Note that, according the previous lemma, not all rank 2 skew-symmetric matrices are TDOA-Matrices. In fact, if $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ in (9) are not coplanar with $\hat{\mathbf{1}}$ (i.e. $(\hat{\mathbf{1}}^\top\hat{\mathbf{v}}_1)\hat{\mathbf{v}}_1 + (\hat{\mathbf{1}}^\top\hat{\mathbf{v}}_2)\hat{\mathbf{v}}_2 \neq \hat{\mathbf{1}}$), then the resulting matrix is rank 2 and skew-symmetric but not a TDOA matrix.

V. TDOA Denoising

In this section we propose a denoising strategy to deal with Gaussian noise in the estimated TDOA measurements, deriving a closed form solution for the proposed optimization problem. This solution is also compared with the Gauss-Markov Estimator.

A. Denoising Strategy

We assume now that each TDOA measurement is contaminated with uncorrelated Gaussian noise $n_{ij} = -n_{ji}$, such that $\Delta\tilde{\tau}_{ij} = \Delta\tau_{ij} + n_{ij}$. Therefore, the measured TDOA matrix $\tilde{\mathbf{M}} = \{\Delta\tilde{\tau}_{ij}\}$ is also a skew-symmetric matrix, sum of a noise-free $\mathbf{M} \in \mathcal{M}_T(n)$ and a skew-symmetric matrix containing noise $\mathbf{N} = \{n_{ij}\}$:

$$\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{N}. \tag{11}$$

Because of the noise, $\tilde{\mathbf{M}} \notin \mathcal{M}_T(n)$ and thus Theorem 1 is no longer satisfied. Consequently, the rank of $\tilde{\mathbf{M}}$ may be higher than two. Nevertheless, we will show that we can take advantage of the structure of TDOA matrices in order to denoise the measured data.

For denoising, we propose finding the closest $\mathbf{M}^* \in \mathcal{M}_T(n)$, to the measured matrix $\tilde{\mathbf{M}}$, in the sense of the Frobenius norm. This approach yields the following optimization problem:

$$\mathbf{M}^* = \underset{\mathbf{M}\in\mathcal{M}_T(n)}{\arg\min} \quad \left\|\tilde{\mathbf{M}} - \mathbf{M}\right\|_F^2. \tag{12}$$

B. Closed-Form Solution

Theorem 3. *Problem* (12) *has the following closed form solution:* $\mathbf{M}^* = (\tilde{\mathbf{M}}\mathbb{1} + \mathbb{1}\tilde{\mathbf{M}})/n$

Proof: From (5), the denoising problem (12) is equivalent to the following constrained convex optimization problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \left\|\tilde{\mathbf{M}} - \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right)\right\|_F^2 \\ \text{subject to} \quad & \hat{\mathbf{1}}^\top \mathbf{x} = 0. \end{aligned} \tag{13}$$

Using the definition of Frobenius norm $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top)$, and trace properties $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ and $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^\top)$ we rewrite the cost as:

$$\begin{aligned} &\left\|\tilde{\mathbf{M}} - \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right)\right\|_F^2 = \\ &= \mathrm{tr}\left(\left[\tilde{\mathbf{M}} - \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right)\right]\left[\tilde{\mathbf{M}} - \left(\mathbf{x}\,\hat{\mathbf{1}}^\top - \hat{\mathbf{1}}\,\mathbf{x}^\top\right)\right]^\top\right) \\ &= 2\left(\mathbf{x}^\top\mathbf{x} - \mathbf{x}^\top\,\hat{\mathbf{1}}\,\hat{\mathbf{1}}^\top\mathbf{x} - \hat{\mathbf{1}}^\top\tilde{\mathbf{M}}^\top\mathbf{x} + \hat{\mathbf{1}}^\top\tilde{\mathbf{M}}\mathbf{x}\right) + \\ &\qquad + \mathrm{tr}\left(\tilde{\mathbf{M}}\tilde{\mathbf{M}}^\top\right) = f\left(\mathbf{x};\tilde{\mathbf{M}}\right). \end{aligned} \tag{14}$$

To solve the constrained problem (13) we use the method of Lagrange multipliers, resulting in the following unconstrained equivalent:

$$\mathbf{x}^* = \underset{\mathbf{x},\lambda}{\arg\min}\left[\Lambda\left(\mathbf{x};\lambda\right)\right], \tag{15}$$

where λ is the Lagrange multiplier and

$$\Lambda\left(\mathbf{x};\lambda\right) = f\left(\mathbf{x};\tilde{\mathbf{M}}\right) + \lambda\hat{\mathbf{1}}^\top\mathbf{x}. \tag{16}$$

We find extrema in (16) by taking first derivatives with respect to both $\mathbf{x}$ and λ and solving the following system:

$$\begin{aligned} &\nabla\Lambda\left(\mathbf{x};\lambda\right) = \hat{\mathbf{0}} \Rightarrow \\ &\begin{cases} 4\mathbf{x}^\top\left(\mathbf{I} - \hat{\mathbf{1}}\,\hat{\mathbf{1}}^\top\right) + 2\hat{\mathbf{1}}^\top\left(\tilde{\mathbf{M}} - \tilde{\mathbf{M}}^\top\right) + \lambda\hat{\mathbf{1}}^\top = \hat{\mathbf{0}}^\top \\ \hat{\mathbf{1}}^\top\mathbf{x} = 0 \end{cases} \end{aligned}$$

$$\begin{aligned} \mathbf{x}^* &= \frac{\left(\tilde{\mathbf{M}} - \tilde{\mathbf{M}}^\top\right)\hat{\mathbf{1}} - \lambda\hat{\mathbf{1}}}{2} \\ \lambda^* &= \frac{\hat{\mathbf{1}}^\top\left(\tilde{\mathbf{M}} - \tilde{\mathbf{M}}^\top\right)\hat{\mathbf{1}}}{2\hat{\mathbf{1}}^\top\hat{\mathbf{1}}}. \end{aligned} \tag{17}$$

Given that the objective function is strictly convex, the solution to the system is unique, and therefore because the proposed solution ($\mathbf{x}^*$ and λ^*) satisfies the equations of a critical point, it is the global minimum

Since $\tilde{\mathbf{M}}$ is skew-symmetric $(\tilde{\mathbf{M}} - \tilde{\mathbf{M}}^\top) = 2\tilde{\mathbf{M}}$. Therefore, (17) becomes:

$$\mathbf{x}^* = \frac{2\tilde{\mathbf{M}}\,\hat{\mathbf{1}} - \lambda\hat{\mathbf{1}}}{2} = \tilde{\mathbf{M}}\,\hat{\mathbf{1}} \tag{18a}$$

$$\lambda = \frac{2\hat{\mathbf{1}}^\top\tilde{\mathbf{M}}\,\hat{\mathbf{1}}}{2\hat{\mathbf{1}}^\top\hat{\mathbf{1}}} = 2\hat{\mathbf{1}}^\top\tilde{\mathbf{M}}\,\hat{\mathbf{1}} = 0. \tag{18b}$$

In (18b) we use the fact that $\hat{\mathbf{1}}^\top\mathbf{A}\,\hat{\mathbf{1}} = 0$ for $\mathbf{A}$ being a skew-symmetric matrix. Also, it is interesting to note from (18a) that $\mathbf{x}^*$ follows the same expression as the one stated in (5) for $\mathbf{x}$ in the noise-free case.

A compact expression for $\mathbf{M}^*$ can be easily derived from (18a) via (5):

$$\begin{aligned} \mathbf{M}^* = \left(\hat{\mathbf{1}}, \tilde{\mathbf{M}}\,\hat{\mathbf{1}}\right)\left(-\tilde{\mathbf{M}}\,\hat{\mathbf{1}}, \hat{\mathbf{1}}\right)^\top = \tilde{\mathbf{M}}\,\hat{\mathbf{1}}\,\hat{\mathbf{1}}^\top + \hat{\mathbf{1}}\,\hat{\mathbf{1}}^\top\tilde{\mathbf{M}} = \\ = (\tilde{\mathbf{M}}\mathbb{1} + \mathbb{1}\tilde{\mathbf{M}})/n. \end{aligned} \tag{19}$$

This completes the proof. ■

Since Lemma 1 relates (5) with SVD, the proposed denoising approach can be considered as a version of Eckart-Young-Mirsky Theorem [34], [35] constrained to TDOA matrices.

C. Equivalence with the Gauss-Markov Estimator

By operating in (19), each element (i, j) of the denoised matrix $\mathbf{M}^*$ is obtained as follows:

$$\mathbf{M}^* = \{\Delta\tau_{ij}^*\} = \left\{\frac{1}{n}\left(\sum_{k=1}^{n}\Delta\tau_{ik} + \Delta\tau_{kj}\right)\right\}. \tag{20}$$

The closed-form in (20) is identical to the one reported in [23, eq.(14)] as the Gauss-Markov estimator of the TDOA measurements, so that all the properties there can be extrapolated to this work. This is not surprising as the least-squares cost of (12) is optimal for Gaussian noise. The same denoising result was found in [25] by projecting TDOA measurements into the set of "feasible" bivectors in a least-squares sense. Under the assumption of Gaussian noise, we can conclude that [23], [25] and our denoising result in (20) are completely equivalent.

VI. Robust TDOA Denoising

In some application scenarios, the assumption of uncorrelated white noise made in section V is fully unrealistic. In cases where the noise is correlated with the signal, measurements are prone to contain outliers in the TDOA measurements due to spurious peaks in the correlation. For such cases, a more complete model for the measured matrix is:

$$\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{N} + \mathbf{S}, \tag{21}$$

where $\mathbf{M} \in \mathcal{M}_T(n)$, $\mathbf{N}$ is a skew-symmetric matrix containing Gaussian noise, much like in (11), and the new matrix $\mathbf{S}$ models the addition of all the outliers. Since the number of outliers is usually small as compared with the number of measurements, we will assume $\mathbf{S}$ to be sparse and unknown.

In order to denoise $\tilde{\mathbf{M}}$, we propose solving the following optimization problem, finding both matrices $\mathbf{M}$ and $\mathbf{S}$:

$$\begin{array}{ll} \underset{\mathbf{M},\mathbf{S}}{\text{minimize}} & \left\|\tilde{\mathbf{M}} - \mathbf{M} - \mathbf{S}\right\|_F^2 \\ \text{subject to} & \mathbf{M} \in \mathcal{M}_T(n) \\ & \|\mathbf{S}\|_0 < 2k, \end{array} \tag{22}$$

where k is the maximum number of outliers supposed to be present in the TDOA measurements.

Robust denoising in (22) is a non-convex optimization problem with constraints that are not even differentiable. This kind of optimization problems have been explored in Robust PCA (RPCA) [36] or robust low-rank factorizations such in GoDec [37]. Despite TDOA matrices are low-rank, these algorithms are not well suited here as they do not include all the algebraic constraints in TDOA matrices.

In order to solve (22), we propose an iterative algorithm, inspired in GoDec. It consists of an alternation method in which $\mathbf{M}$ and $\mathbf{S}$ are obtained in turns, with close-form solutions for these two steps (we use a subindex t to denote the iteration count):

$$\begin{cases} \mathbf{M}_t = \underset{\mathbf{M}\in\mathcal{M}_T(n)}{\arg\min} & \left\|\tilde{\mathbf{M}} - \mathbf{M} - \mathbf{S}_{t-1}\right\|_F^2 \\ \mathbf{S}_t = \underset{\|\mathbf{S}\|_0<2k}{\arg\min} & \left\|\tilde{\mathbf{M}} - \mathbf{M}_t - \mathbf{S}\right\|_F^2 \end{cases} \tag{23}$$

The first sub-problem of (23) is the same as our denoising problem in (12), therefore $\mathbf{M}_t$ can be updated via (19). Then, $\mathbf{S}_t$ is updated via entry-wise hard thresholding of $\tilde{\mathbf{M}} - \mathbf{M}_t$. Thus:

$$\begin{cases} \mathbf{M}_t = \left(\tilde{\mathbf{M}} - \mathbf{S}_{t-1}\right)\hat{\mathbf{1}}\hat{\mathbf{1}}^\top + \hat{\mathbf{1}}\hat{\mathbf{1}}^\top\left(\tilde{\mathbf{M}} - \mathbf{S}_{t-1}\right) \\ \mathbf{S}_t = \mathcal{P}_{2k}\left(\tilde{\mathbf{M}} - \mathbf{M}_t\right) \end{cases} \tag{24}$$

where $\mathcal{P}_l(\mathbf{X})$ is an function which generates a matrix with the same size of $\mathbf{X}$, preserving the l elements of $\mathbf{X}$ with the largest absolute value, and making the rest of elements zero. Note that, since $\mathbf{X}$ is skew symmetric in our application, the result provided by $\mathcal{P}_{2k}(\cdot)$ is also skew symmetric. The convergence to a local minimum of this algorithm is guaranteed in similar circumstances as GoDec [37], as the solutions to both sub-problems in (24) are solved globally.

So, the proposed robust denoising algorithm is shown in Alg. 1.

Algorithm 1 Robust denoising.

Require: $\tilde{\mathbf{M}}$, k, ϵ
Ensure: $\mathbf{M} \in \mathcal{M}_T(n)$, $\|\mathbf{S}\|_0 < 2k$,
1: $\mathbf{M}_0 = 0$; $\mathbf{S}_0 = 0$; $t = 0$
2: **while** $\|\tilde{\mathbf{M}} - \mathbf{M}_t - \mathbf{S}_t\|_F^2 / \|\tilde{\mathbf{M}}\|_F^2 > \epsilon$ **do**
3: $\quad t = t + 1$
4: $\quad \mathbf{M}_t = (\tilde{\mathbf{M}} - \mathbf{S}_{t-1})\hat{\mathbf{1}}\hat{\mathbf{1}}^\top + \hat{\mathbf{1}}\hat{\mathbf{1}}^\top(\tilde{\mathbf{M}} - \mathbf{S}_{t-1})$
5: $\quad \mathbf{S}_t = \mathcal{P}_{2k}(\tilde{\mathbf{M}} - \mathbf{M}_t)$
6: **end while**
7: **return** $\mathbf{M}_t$, $\mathbf{S}_t$

From now on, we will refer to this algorithm as `Robust DeN`.

VII. Missing Data Recovery

A. Recovery Strategy

In real scenarios, there may be situations where some of the elements of $\tilde{\mathbf{M}}$ might not be available (for instance, due to communications failure) or even when they are available, there are reasons to avoid using them (for example, due to a priori knowledge of unreliable measurements, or when calculating the whole redundant set is computationally too demanding) [32]. In such cases, we want to be able to avoid some measurements, thus performing estimations when part of the values in $\tilde{\mathbf{M}}$ are missing.

In this section, we address the matrix completion problem ([38], [39]) for TDOA matrices. We assume that in a measured TDOA matrix $\tilde{\mathbf{M}}$, some of its elements are unknown, and the rest are contaminated with additive Gaussian noise. We take advantage of the redundancy present in TDOA matrices to estimate a complete denoised TDOA matrix including the missing entries.

The matrix completion problem is stated as follows:

$$\mathbf{M}^* = \underset{\mathbf{M}\in\mathcal{M}_T(n)}{\arg\min} \quad \left\|\mathbf{L} \circ \left(\tilde{\mathbf{M}} - \mathbf{M}\right)\right\|_F^2, \tag{25}$$

where $\mathbf{L}$ is a symmetric binary matrix whose element (i, j) is 1 if the TDOA between the sensor i and j is known, being 0 otherwise. For convenience and without loss of generality, the elements on the main diagonal of $\mathbf{L}$ will be set to 1.

Solving (25) is equivalent to finding the full TDOA matrix whose elements best fit the available elements of $\tilde{\mathbf{M}}$. Note that, $\mathbf{L} \circ (\tilde{\mathbf{M}} - \mathbf{M}) = (\tilde{\mathbf{M}}_\mathbf{L} - \mathbf{L} \circ \mathbf{M})$, where $\tilde{\mathbf{M}}_\mathbf{L} = (\mathbf{L} \circ \tilde{\mathbf{M}})$ is the result of setting the unknown elements of $\tilde{\mathbf{M}}$ to zero.

B. Closed-Form Solution

Theorem 4. *The problem* (25) *has the following closed form solution:* $\mathbf{M}^* = \left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)^{-1}\tilde{\mathbf{M}}_\mathbf{L}\mathbb{1} + \mathbb{1}\tilde{\mathbf{M}}_\mathbf{L}\left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)^{-1}$ *where* $\mathbf{D}_\beta = \left(\mathbf{I} \circ \mathbf{L}\mathbf{L}^\top\right)$ *is a* $n \times n$ *diagonal matrix with* $\boldsymbol{\beta} = \left(n - \bar{\beta}_1, \cdots, n - \bar{\beta}_n\right)^\top = \sqrt{n}\,\mathbf{L}\,\hat{\mathbf{1}}$ *as its main diagonal.* $\bar{\beta}_i$ *is the number of missing measurements with the sensor* i.

Proof: Using Corollary 2.2, problem (25) is rewritten as

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \left\|\tilde{\mathbf{M}}_\mathbf{L} - \dfrac{\mathbf{L} \circ (\mathbf{D}_\mathbf{x}\mathbb{1} - \mathbb{1}\mathbf{D}_\mathbf{x})}{\sqrt{n}}\right\|_F^2 \\ \text{subject to} & \hat{\mathbf{1}}^\top\mathbf{x} = 0. \end{array} \tag{26}$$

Since $\mathbb{1}$ is the identity element of the hadamard product and $\mathbf{D_x}$ is a diagonal matrix, we can rewrite (26) as:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \left\| \tilde{\mathbf{M}}_{\mathbf{L}} - \frac{(\mathbf{D_x L} - \mathbf{L D_x})}{\sqrt{n}} \right\|_F^2 \\ \text{subject to} \quad & \hat{\mathbf{1}}^\top \mathbf{x} = 0. \end{aligned} \tag{27}$$

Operating in a similar manner to (14) we get:

$$\begin{aligned} \left\| \tilde{\mathbf{M}}_{\mathbf{L}} - \frac{(\mathbf{D_x L} - \mathbf{L D_x})}{\sqrt{n}} \right\|_F^2 &= \frac{2}{n} \operatorname{tr}\left(\mathbf{D_x L L}^\top \mathbf{D_x}\right) - \\ - \frac{2}{n} \operatorname{tr}\left(\mathbf{D_x L D_x L}\right) + \frac{2}{\sqrt{n}} \operatorname{tr}\left(\left[\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right] \mathbf{D_x L}\right) + \\ + \operatorname{tr}\left(\tilde{\mathbf{M}}_{\mathbf{L}} \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right). \end{aligned} \tag{28}$$

Using the identity $\mathbf{x}^* \left(\mathbf{A} \circ \mathbf{B}\right) \mathbf{y} = \operatorname{tr}\left(\mathbf{D_x}^* \mathbf{A D_y B}^\top\right)$ we get:

$$\begin{aligned} \left\| \tilde{\mathbf{M}}_{\mathbf{L}} - \frac{(\mathbf{D_x L} - \mathbf{L D_x})}{\sqrt{n}} \right\|_F^2 &= \frac{2}{n} \mathbf{x}^\top \left(\mathbf{I} \circ \mathbf{L L}^\top\right) \mathbf{x} \\ - \frac{2}{n} \mathbf{x}^\top \left(\mathbf{L} \circ \mathbf{L}^\top\right) \mathbf{x} + 2\, \hat{\mathbf{1}}^\top \left(\left[\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right] \circ \mathbf{L}^\top\right) \mathbf{x} + \\ + \operatorname{tr}\left(\tilde{\mathbf{M}}_{\mathbf{L}} \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) = g\left(\mathbf{x}; \tilde{\mathbf{M}}, \mathbf{L}\right) \end{aligned} \tag{29}$$

and finally:

$$\begin{aligned} \left\| \tilde{\mathbf{M}}_{\mathbf{L}} - \frac{(\mathbf{D_x L} - \mathbf{L D_x})}{\sqrt{n}} \right\|_F^2 = \\ \frac{2}{n} \left(\mathbf{x}^\top \mathbf{D}_\beta \mathbf{x} - \mathbf{x}^\top \mathbf{L} \mathbf{x} + n\, \hat{\mathbf{1}}^\top \left(\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) \mathbf{x}\right) + \\ + \operatorname{tr}\left(\tilde{\mathbf{M}}_{\mathbf{L}} \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) = g(\mathbf{x}; \tilde{\mathbf{M}}, \mathbf{L}). \end{aligned} \tag{30}$$

It is important to note that equations (14) and (30) are identical when there is no missing data in $\tilde{\mathbf{M}}$ (i.e $\mathbf{L} = \mathbb{1} = n\, \hat{\mathbf{1}}\, \hat{\mathbf{1}}^\top$ and $\mathbf{D}_\beta = n\, \mathbf{I}$).

We use the method of Lagrange multipliers to express (26) as the following unconstrained optimization problem:

$$\Lambda\left(\mathbf{x}; \lambda\right) = g\left(\mathbf{x}; \tilde{\mathbf{M}}, \mathbf{L}\right) + \lambda \hat{\mathbf{1}}^\top \mathbf{x}. \tag{31}$$

By taking derivatives we obtain the following system:

$$\begin{aligned} & \nabla \Lambda\left(\mathbf{x}; \lambda\right) = \hat{\mathbf{0}} \Rightarrow \\ & \begin{cases} \frac{2}{n} \mathbf{x}^\top \left(\mathbf{D}_\beta - \mathbf{L}\right) + \hat{\mathbf{1}}^\top \left(\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) + \lambda \hat{\mathbf{1}}^\top = \hat{\mathbf{0}} \\ \hat{\mathbf{1}}^\top \mathbf{x} = 0. \end{cases} \end{aligned} \tag{32}$$

Since $\hat{\mathbf{1}}^\top \mathbf{x} = 0$ implies that $\mathbb{1}\mathbf{x} = \hat{\mathbf{0}}$, we substitute in (32) obtaining:

$$\frac{2}{n} \left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right) \mathbf{x} = \left(\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) \hat{\mathbf{1}} - \lambda \hat{\mathbf{1}}, \tag{33}$$

where $\left(\bar{\mathbf{L}} = \mathbb{1} - \mathbf{L}\right)$ (logical not operator over all elements of $\mathbf{L}$). Note that: $\left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)$ is symmetric and, furthermore, $\hat{\mathbf{1}}$ is one of its eigenvectors.

$$\left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right) \hat{\mathbf{1}} = \frac{\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}}{\sqrt{n}} = n \hat{\mathbf{1}}. \tag{34}$$

Therefore, if the two terms of (33) are multiplied on the right by $\hat{\mathbf{1}}^\top$ we get:

$$2\, \hat{\mathbf{1}}^\top \mathbf{x} = \hat{\mathbf{1}}^\top \left(\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top\right) \hat{\mathbf{1}} - \lambda. \tag{35}$$

Then applying to (35) the fact that

$$\hat{\mathbf{1}}^\top \mathbf{x} = 0 \quad \text{and} \quad \hat{\mathbf{1}}^\top (\tilde{\mathbf{M}}_{\mathbf{L}} - \tilde{\mathbf{M}}_{\mathbf{L}}^\top) \hat{\mathbf{1}} = 0, \tag{36}$$

we can conclude that $\lambda = 0$. Thus, the solution of (26) is:

$$\mathbf{x}^* = n \left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)^{-1} \tilde{\mathbf{M}}_{\mathbf{L}} \hat{\mathbf{1}}. \tag{37}$$

Finally, the solution of problem (25) can be calculated from (37) using (5):

$$\mathbf{M}^* = \left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)^{-1} \tilde{\mathbf{M}}_{\mathbf{L}} \mathbb{1} + \mathbb{1} \tilde{\mathbf{M}}_{\mathbf{L}} \left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)^{-1} \tag{38}$$

This completes the proof. ■

From now on, we will refer to this algorithm as `MC`. It is noteworthy to comment that the matrix $\left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)$ contains important information about the recoverability of missing data: if it is full-rank, then the solution of (25) is unique and if $\left(\mathbf{D}_\beta + \bar{\mathbf{L}}\right)$ is rank-deficient, missing data is not recoverable uniquely without any further assumption.

Furthermore, in the absence of missing data, $n(\mathbf{D}_\beta + \bar{\mathbf{L}})^{-1} = \mathbf{I}$, hence the matrix completion solution in (38) becomes the solution of the denoising problem stated in (19).

VIII. Robust TDOA Denoising with Missing Data

In this section we aim to combine the results of sections VI and VII, addressing the more general case in which both outliers and missing data are considered. Therefore, the problem is a combination of (22) and (25) defined as:

$$\begin{aligned} \underset{\mathbf{M},\mathbf{S}}{\text{minimize}} \quad & \left\| \mathbf{L} \circ \left(\tilde{\mathbf{M}} - \mathbf{M} - \mathbf{S}\right) \right\|_F^2 \\ \text{subject to} \quad & \mathbf{M} \in \mathcal{M}_T(n) \\ & \|\mathbf{S}\|_0 < 2k \\ & \mathbf{S} = \mathbf{L} \circ \mathbf{S}. \end{aligned} \tag{39}$$

In the same way as in section VI, (39) can be solved by alternatively solving the following two subproblems until convergence:

$$\mathbf{M}_t = \underset{\mathbf{M} \in \mathcal{M}_T(n)}{\arg\min} \quad \left\| \mathbf{L} \circ \left(\tilde{\mathbf{M}} - \mathbf{M} - \mathbf{S}_{t-1}\right) \right\|_F^2 \tag{40a}$$

$$\mathbf{S}_t = \underset{\|\mathbf{S}\|_0 < 2k}{\arg\min} \quad \left\| \mathbf{L} \circ \left(\tilde{\mathbf{M}} - \mathbf{M}_t\right) - \mathbf{S} \right\|_F^2. \tag{40b}$$

The subproblem (40a) is equivalent to the missing data problem solved in section VII but considering $\tilde{\mathbf{M}}_{\mathbf{L}} = (\mathbf{L} \circ \tilde{\mathbf{M}} - \mathbf{S}_{t-1})$. Therefore, according to Theorem 4, it has a closed form solution:

$$\begin{aligned} \mathbf{M}_t^* = (\mathbf{D}_\beta + \bar{\mathbf{L}})^{-1} (\mathbf{L} \circ \tilde{\mathbf{M}} - \mathbf{S}_{t-1})\, \mathbb{1} + \\ + \mathbb{1} (\mathbf{L} \circ \tilde{\mathbf{M}} - \mathbf{S}_{t-1}) (\mathbf{D}_\beta + \bar{\mathbf{L}})^{-1}. \end{aligned} \tag{41}$$

Since (40b) is of the same form as the second subproblem in (23), it can also be solved by entry-wise hard thresholding of $\mathbf{L} \circ (\tilde{\mathbf{M}} - \mathbf{M}_t)$.

The pseudocode shown in Alg. 2 summarizes the proposed algorithm for the general case.

Algorithm 2 Robust denoising with missing data.

Require: $\tilde{\mathbf{M}}$, $\mathbf{L}$, k, ϵ
Ensure: $\mathbf{M} \in \mathcal{M}_T(n)$, $\|\mathbf{S}\|_0 < 2k$,
1: $\mathbf{D}_\beta = \mathbf{I} \circ \mathbf{L}\mathbf{L}^\top$
2: $\mathbf{Q} = (\mathbf{D}_\beta + \bar{\mathbf{L}})^{-1}$
3: $\mathbf{M}_0 = 0$; $\mathbf{S}_0 = 0$; $t = 0$
4: **while** $\|\tilde{\mathbf{M}} - \mathbf{M}_t - \mathbf{S}_t\|_F^2 / \|\tilde{\mathbf{M}}\|_F^2 > \epsilon$ **do**
5: $\quad t = t + 1$
6: $\quad \mathbf{M}_t = \mathbf{Q}(\mathbf{L} \circ \tilde{\mathbf{M}} - \mathbf{S}_{t-1})\mathbb{1} + \mathbb{1}(\mathbf{L} \circ \tilde{\mathbf{M}} - \mathbf{S}_{t-1})\mathbf{Q}$
7: $\quad \mathbf{S}_t = \mathcal{P}_{2k}(\tilde{\mathbf{M}} - \mathbf{M}_t)$
8: **end while**
9: **return** $\mathbf{M}_t$, $\mathbf{S}_t$

Note that in line 2 the matrix $\mathbf{Q} = (\mathbf{D}_\beta + \bar{\mathbf{L}})^{-1}$ can be precalculated in order to get an efficient implementation of the algorithm.

From now on, we will refer to this algorithm as `Robust DeN+MC`.

IX. Experiments with Synthetic Data

In this section computer simulations will be used to compare the proposed algorithms with some of the alternatives existing in the state of the art.

For evaluating the `Robust DeN` and `Robust DeN+MC` algorithms, two different metrics will be used:

- The Signal-to-Noise-Ratio SNR [dB] of the *non-redundant set* referenced to the first sensor ($10\log(\sum_{i=1}^{n}\|\Delta\tau_{i1}\|^2 / \sum_{i=1}^{n}\|\Delta\tau_{i1}^* - \Delta\tau_{i1}\|^2)$). This is an application independent metric (where $\Delta\tau_{i1}^*$ is the estimation of $\Delta\tau_{i1}$), that will allow assessing the proposal improvements in the TDOA measurements *per se*.
- The localization error, measured as the average distance between the source ground truth position and the position estimated using any given localization algorithm based on TDOA estimations (such as [16] in our case). This is an application dependent metric, that will allow assessing the actual benefits of the proposal in an example of a real task. Note that our proposal is not restricted to localization and can be used in other applications that could benefit from denoised TDOAs (such as self-calibration or beamforming).

A. Experimental setup

For all the synthetic data experiments, a set of 10 sensors (which implies 45 different sensor pairs) and 1 source were randomly located. Therefore, 45 different TDOA measurements were generated per experiment, and independent Gaussian noise was added to them, using the same variance for all the measurements.

The sensor locations were uniformly distributed in a cube of 1 meter side, and the source positions were uniformly distributed in a 2 meter side cube. The propagation speed of the signal was set to 343.313 m/s. In all the experiments where it's required, ϵ is set to 10^{-10}. To increase the statistical significance of the results, they are provided as averages of 20 independent runs.

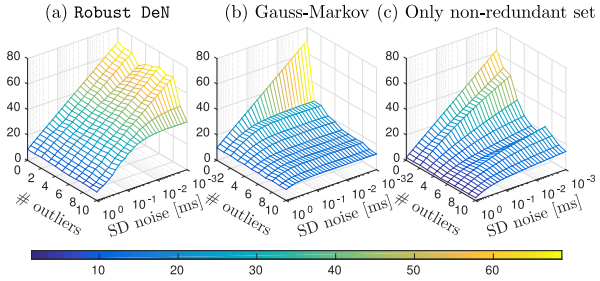


Fig. 2: Robust denoising in synthetic data: SNR in dB, higher is better.

B. Evaluation of Robust TDOA Denoising

In this first experiment, we evaluated the performance of the `Robust DeN` algorithm proposed in section VI, imposing that some TDOA values were outliers. To simulate this, we randomly chose some measurements (between 0 and 10) and replaced them with a zero-mean Gaussian distributed noise, with a standard deviation of 0.1 ms. It is worth mentioning that the outlier values calculated that way are not related at all to the real TDOAs, thus being *true* outliers. The parameter k of the proposed algorithm, which fixes the maximum number of identifiable outliers, was set to 8.

1) SNR Improvements Evaluation: Fig. 2a shows the SNR values for the `Robust DeN` algorithm when modifying the noise standard deviation and the number of outliers, compared with that obtained by the Gauss-Markov estimator (Fig. 2b), and also when only the *non-redundant set* is used, i.e. not using the redundancy of TDOA measurements (Fig. 2c).

As predicted in section V, when no outliers are present, the performance of the `Robust DeN` algorithm is the same as Gauss-Markov (see row 0 in Figs. 2a and 2b), hence it reaches the Cramer-Rao Bound [23], while being much better than using no redundancy. Nevertheless, the proposed algorithm clearly outperforms the other two approaches when outliers are present in the measurements (rows 1 through 10 in the graphics of Fig. 2).

2) Source Localization Improvements Evaluation: The optimized *non-redundant set* provided by the algorithms applied in Section IX-B1 were used in a localization algorithm using [16]. The average localization errors (in mm) are shown in Fig. 3. Again, the `Robust DeN` algorithm performs as Gauss-Markov when there are no outliers, but is clearly superior when outliers are present.

It is also worth mentioning that the behaviour of the robust denoising keeps the improvements at roughly the same level for increasing number of outliers present, thus validating the ability of the algorithm to pinpoint and eliminate their presence.

C. Evaluation of Missing Data Recovery

In this second experiment, we evaluated the capability of the `MC` algorithm proposed in section VII to recover missing values. For our purposes, the missing TDOA measurements were also chosen randomly but, in contrast to the previous experiment, the matrix positions of the missing measurements were known.

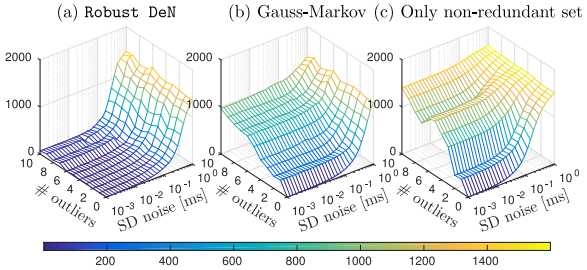


Fig. 3: Robust denoising in synthetic data: Localization error in mm (using [16]), lower is better.

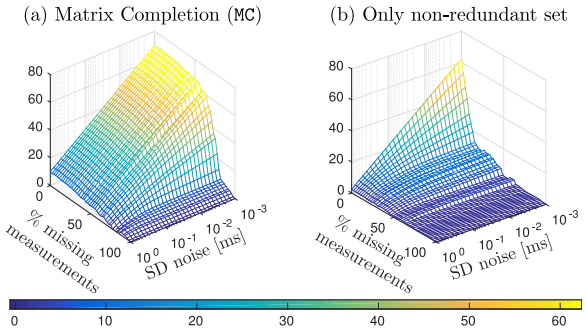


Fig. 4: Missing data recovery in synthetic data: SNR in dB, higher is better.

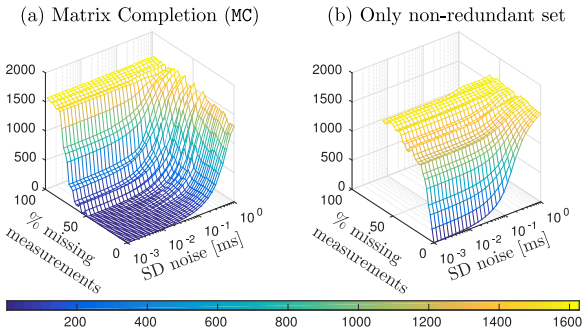


Fig. 5: Missing data recovery in synthetic data: Localization error in mm (using [16]), lower is better.

Fig. 4 and Fig. 5 show, respectively, the SNR values, and the localization error for the MC algorithm, when modifying the noise standard deviation and the percentage of missing TDOA values in the TDOA matrix, as compared with using only the *non-redundant set* (missing values were set to zero).

From the figures, it can be clearly seen that the proposed algorithm can take advantage of the knowledge about which measurements were missing, achieving even better results than when the positions of the outliers were unknown. For example, removing 50% of the full TDOA set of measurements implies 23 missing values of 45 measurements , much more than the maximum of 10 outliers evaluated in Fig. 2, while keeping good performance.

D. Evaluation of Robust TDOA Denoising with Missing Data

In this third experiment, we evaluated the capability of the Robust DeN+MC algorithm proposed in section VIII to face both outliers and recover missing values.

To provide a wide range of evaluation scenarios, we defined: *i)* Two conditions related to noise, namely *low* and *high*. The former corresponds to a standard deviation of 10^{-3} ms., and the latter to 0.2 ms. *ii)* Two conditions related to the presence of outliers, imposing the existence of 2 or 6 outliers. *iii)* A variable number of missing TDOA measurements, defined as a percentage of missing TDOA values in the TDOA matrix.

In all cases, the number of outliers is fixed among the full TDOA set and then, some measurements are discarded, i.e. some of the discarded measurements may be outliers. Note that this is consistent with the real case, where it is not possible to anticipate where the outliers are.

Fig. 6 and Fig. 7 show, respectively, the SNR values, and the localization error for different algorithms, and for different evaluation scenarios.

As it can be seen in Fig. 6a, 6c, 7a, and 7c, when there are a low number of outliers (2 in this case), the best results are obtained for lower k values. However, when the number of outliers increase, (Fig. 6b, 6d, 7b and 7d, low k values perform worse. So, we can conclude that k must be a number as low as possible, but higher than the number of actual outliers.

Nevertheless, it is worth to observe the behaviour of Fig. 6b, 6d, 7b and 7d (with more outliers) when the percentage of missing data increases. It can be clearly seen that the lines corresponding to different values of k are crossing among them. This seems to indicate that as the missing data percentage increases, the number of outliers that we are able to detect decreases.

Anyway, the results obtained by the Robust DeN+MC algorithm outperforms the Gauss-Markov estimator, asymptotically approaching it when the noise is very high. Note also that for high values of noise, the noise and the outliers are practically indistinguishable.

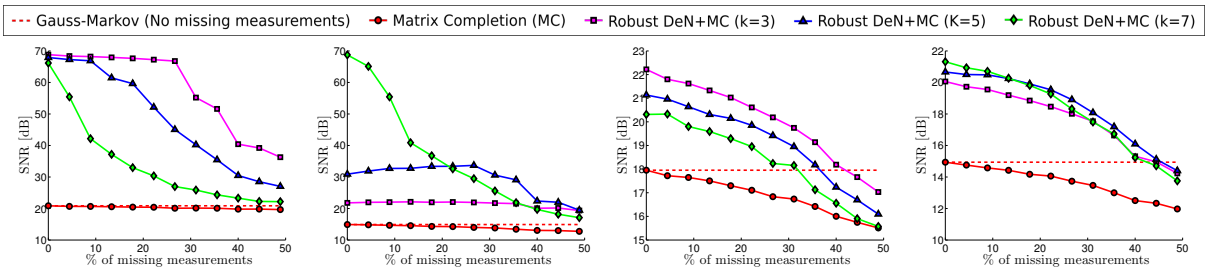
X. Experiments with Real Data

The aim of this section is to evaluate whether the improvements obtained in section IX using synthetic data are actually found in real environments. To do this, the proposed algorithms have been evaluated using audio recordings from the AV16.3 database [40], an audio-visual corpus recorded in the *Smart Meeting Room* of the IDIAP Research Institute, in Switzerland. The same metrics and localization algorithm of the previous section has been employed.

Additionally, the proposed Robust DeN algorithm is compared with other recent state-of-the-art methods in section X-D. In that case, we have employed the same framework used in [8], wherein many methods were already compared. In order to make the comparison as fair as possible, in this part the localization algorithm will be the same as in [8].

A. Experimental Setup

The IDIAP Meeting Room (shown in Fig. 8) is a $8.2m \times 3.6m \times 2.4m$ rectangular space containing a centrally



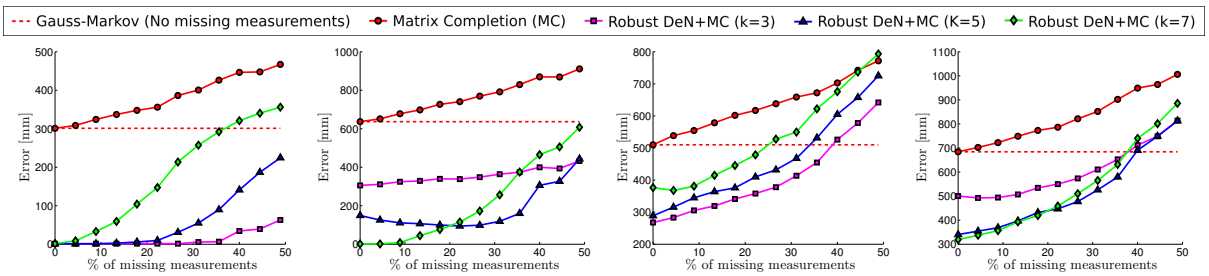
(a) SNR with low noise and 2 outliers

(b) SNR with low noise and 6 outliers

(c) SNR with high noise and 2 outliers

(d) SNR with high noise and 6 outliers

Fig. 6: Algorithm evaluation in synthetic data: SNR in dB.



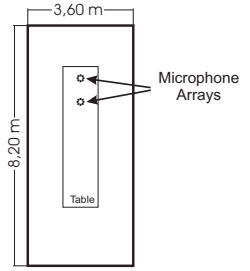
(a) Localization error with low noise and 2 outliers

(b) Localization error with low noise and 6 outliers

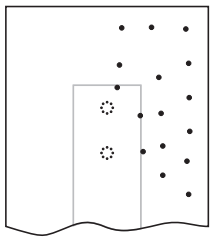
(c) Localization error with high noise and 2 outliers

(d) Localization error with high noise and 6 outliers

Fig. 7: Algorithm evaluation in synthetic data: Localization error in mm ([16] is used for source position estimation from the optimized TDOA values).



(a) IDIAP room layout showing the centered table, and the microphones arranged in two circular arrays.



(b) Positions evaluated in the real data experiments. Only the relevant room section is shown.

Fig. 8: IDIAP Smart Meeting Room: experimental details.

located $4.8m \times 1.2m$ rectangular table, on top of which two circular microphone arrays of $10cm$ radius are located, each of them composed by 8 microphones. The centers of the two arrays are separated by $80cm$ and the origin of coordinates is located in the middle point between the two arrays. A detailed description of the meeting room can be found in [41].

The audio recordings are synchronously sampled at 16 KHz, and the complete database along with the corresponding annotation files containing the recordings ground truth (3D coordinates of the speaker's mouth) is fully accessible on-line at [42]. It is composed by several sequences from which we are using sequence 01, with a single male speaker generating digit strings in 16 positions (which can be seen as small circles in Fig. 8b), distributed along the room. The sequence duration accounts for 208 seconds in total, with 823 ground truth frames.

The TDOA measurements $\Delta\tilde{\tau}_{ij}$, from which the measured TDOA matrix $\tilde{\mathbf{M}}$ is built, where estimated using the highest peak of the GCC-PHAT function[26].

As in a real scenario outliers are common but difficult to anticipate or enforce, the sweep over noise levels and the number of outliers that we performed with synthetic data are not feasible. Therefore, in our experiments with real data, we will only provide the SNR values and localization errors obtained after using each algorithm.

B. Evaluation of Robust TDOA Denoising

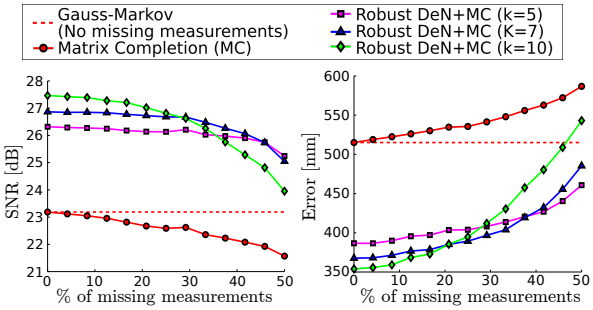
In this experiment, *all* the microphone pairs have been considered, hence 120 TDOA values have been computed for each frame.

In table I we show an example of the results for the `Robust DeN` algorithm with $k = 10$. As it also happened with synthetic data, in this case the proposed algorithm outperforms the Gauss-Markov estimator, yielding great improvements in both SNR and localization precision. Fig. 9 shows that the selection of k is not very critical in this dataset as improvements over Gauss-Markov are obtained for a wide range of k.

These results are the baseline for the experiments with missing data described in the next subsection.

TABLE I: Robust denoising performance in real data

	SNR (dB)	Average Localization error (mm)
`Robust DeN`	27.46	354
Gauss-Markov	23.19	515
Only non-redundant set	17.83	858



(a) SNR in dB

(b) Localization error in mm (using [16])

Fig. 9: Results for real data with missing TDOA measurements.

C. Evaluation of Robust TDOA Denoising with Missing Data

In the second experiment with real data, we randomly remove a set of TDOA measurements. Fig. 9 shows the obtained results. The dotted lines correspond to the performance (SNR and localization error) achieved by the Gauss-Markov estimator when there are no missing measurements. The solid lines with circular marks are the results obtained by the `MC` algorithm described in section VII.

On the other hand, the solid lines with squared/triangular/diamond marks correspond with the results of the `Robust DeN+MC` algorithm presented in section VIII. The different colors/shapes indicate different values of the hyperparameter k.

Fig. 9 highlights the relevance of the proposed robust-denoising algorithm (`Robust DeN+MC`) in real-scenarios, with important improvements over its non-robust version (`MC`): higher than 4dB absolute improvement in terms of SNR, and around 30% relative improvement (15 cm absolute) in terms of localization precision.

We again observe that as the percentage of missing data increases, the lines corresponding to different values of k are crossing among them. This behaviour is very similar to that found in the synthetic experiments above (refer to Fig. 6 and Fig. 7) for a high number of outliers, what suggests that this is the case in the real experiment, also serving as a validation for our simulation conclusions.

It is also noteworthy that, in order to get the best result, the maximum number of outliers k should be decreased when the number of missing measurements increases.

D. Comparison with Other Methods in a Localization Task

In this section we made use of the code and real data of the gtde MATLAB toolbox [8], [43]. In this toolbox, real recordings were performed under noisy conditions in a 4x4x4 (m) room, in which an array of 4 microphones forming a tetrahedron of 20 cm side has been placed. Sound waves coming from a loudspeaker placed at 189 different locations 1.7 m away from the array were recorded at 48 KHz.

Table II shows the results of the different algorithms on the localization task. The first 5 columns refer to the results of the `Robust DeN` algorithm for 5 different values of the parameter k. The rest of the columns show the results of a selection of the algorithms implemented and evaluated in [8] (the nomenclature has been kept and the results are essentially the same).

From [8], we selected three multilateration methods (generically referred to as *x-mult*) which are implementations of the algorithm described in [44]. These methods require to be initialized with the distance r to the source[1]. They were selected because they also make use of the redundancy between all the correlations, using the same input as the `Robust DeN` algorithm. We also compare our proposal with the branch & bound (*b&b*) method proposed by the authors of [8], as this is the one with the best performance in that work. The interested reader may refer to [8] for further details about the used methods, and the results of some other methods as well.

The `Robust DeN` algorithm is the only one of the evaluated algorithms that does not require information about the array geometry to perform denoising of the TDOA estimations. The localization algorithm (i.e converting TDOA to angles), implemented in [43], was used after all the methods in order to make the comparison as fair as possible.

The first row in table II shows the percentage of localization measurements with an angular error lower than 30º, the second row shows their mean angular error, and the third row their standard deviation.

To complement the data of Table II, we have also evaluated the execution timing details of the evaluated algorithms, with the results shown in Table III.

Table II shows that the `Robust DeN` algorithm performs better than the *x-mult* algorithms. Note that this happens even when the input data is the same, and neither the geometry of the array, nor the distance to the source r are used for denoising in our proposal. In what respect to computational demands, our proposal is over 50 times faster.

Comparing with the *b&b* algorithm, our proposal is close to its performance, which is also an important result provided that (again) we do not use the array geometry for denoising, and our execution time is over 110 times faster.

XI. Conclusions

This paper has studied the properties of TDOA matrices, showing that they can be effectively used for solving TDOA denoising problems. In particular, the paper has investigated challenging scenarios where the TDOA matrix is contaminated with Gaussian noise, outliers and where a percentage of the measurements are missing. The paper shows that denoising in the presence of Gaussian noise and missing data can be solved in closed-form. This result is important, as it is the basis of an iterative algorithm that can also cope with

[1] *n-mult*, *t-mult* and *f-mult*, use r values of $0.9m$, $1.7m$ and $2.5m$ respectively, following [8].

TABLE II: Real data performance comparison between `Robust DeN` and selected algorithms in [8] (marked with *) ($T_{60} \approx 0.5s$)

	Robust DeN					*n-mult**	*t-mult**	*f-mult**	*b&b**
	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$				
% Measurements with angular error $< 30^\circ$	23.78%	24.12%	25.20%	25.71%	23.52%	13.99%	17.38%	17.91%	27.60%
Mean angular error	17.21	16.44	16.60	16.78	17.66	17.08	16.05	15.25	16.89
Standard deviation of angular error	7.52	7.44	7.52	7.52	7.38	7.31	7.80	7.78	7.50

TABLE III: Average execution time (s) and standard deviation for each method (evaluated on 9435 trials).

	`Robust DeN`	*x-mult*	*b&b*
Mean time (s)	0.024	1.255	2.649
Std (s)	0.0011	0.086	0.339

outliers. The paper has tested the proposed algorithms in the context of acoustic localization using microphone arrays. The experimental results, both on real and synthetic data have shown that our algorithms successfully perform denoising (up to 30% of improvement in localization accuracy) with a high rate of missing data (up to 50%) and outliers, without knowing the sensor positions. This is important as it opens its application to tasks where the sensors geometry is unknown. Interestingly, in real datasets our robust denoising algorithm is systematically better than the Gauss-Markov estimator even when there is no missing data. This is also an important result as it proves that the assumption of Gaussian noise does not hold in real cases, while our robust model is capable of automatically discard erroneous measurements. The proposed robust denoising method has also been compared with other methods in the literature on a localization task. Our results are very similar to the state of the art, even though we do not require knowing the array geometry in the denoising stage. Furthermore, our proposal is significantly less computationally demanding.

As for future work, we plan to further test our denoising algorithms in applications where the position of the sensors is unknown in advance, such in self-localization and beamforming.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R), and by the University of Alcalá under projects DETECTOR and ARMIS. Afsaneh Asaei is supported by funding from SNSF project PHASER, grant agreement number 200021-153507. We specially thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

REFERENCES

[1] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44 – 53, jan. 2005.

[2] Y. Kuang and K. Astrom, "Stratified sensor network self-calibration from TDOA measurements," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[3] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[4] M. Brandstein and D. Ward, Eds., *Microphone Arrays : Signal Processing Techniques and Applications*. Springer, 2001.

[5] G. C. Carter *et al.*, "Special issue on time delay estimation," in *IEEE Trans. Acoust., Speech, Sig. Process.*, G. C. Carter, Ed., Jun 1981, vol. ASSP-29, no. 3, pp. 461 – 624.

[6] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 170–170, 2006.

[7] K. Ho, "Bias reduction for an explicit solution of source localization using TDOA," *Signal Processing, IEEE Transactions on*, vol. 60, no. 5, pp. 2101–2114, 2012.

[8] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 6, pp. 1082–1095, 2014.

[9] M. Compagnoni, R. Notari, F. Antonacci, and A. Sarti, "A comprehensive analysis of the geometry of TDOA maps in localization problems," *Inverse Problems*, vol. 30, no. 3, p. 035004, 2014.

[10] A. Nouvellet, M. Charbit, F. Roueff, and A. Le Pichon, "Slowness estimation from noisy time delays observed on non-planar arrays," *Geophysical Journal International*, vol. 198, no. 2, pp. 1199–1207, 2014.

[11] B. Huang, L. Xie, and Z. Yang, "TDOA-based source localization with distance-dependent noises," *Wireless Communications, IEEE Transactions on*, vol. 14, no. 1, pp. 468–480, 2015.

[12] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-04619-7_8

[13] B. Yang and J. Scheuing, "A theoretical analysis of 2D sensor arrays for TDOA based localization," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4. IEEE, 2006, pp. IV–IV.

[14] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *Information Theory, IEEE Transactions on*, vol. 19, no. 5, pp. 608–614, Sep 1973.

[15] N. Zhu, "Locating and extracting acoustic and neural signals," Wayne State University, Dissertation 422, 2011. [Online]. Available: http://digitalcommons.wayne.edu/oa_dissertations/422

[16] Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *Signal Processing, IEEE Transactions on*, vol. 42, no. 8, pp. 1905–1915, 1994.

[17] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data." in *ICASSP*, 2008, pp. 2445–2448.

[18] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 12, pp. 1661–1669, 1987.

[19] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," *Signal Processing Letters, IEEE*, vol. 15, pp. 1–4, 2008.

[20] Y. Weng, W. Xiao, and L. Xie, "Total least squares method for robust source localization in sensor networks using TDOA measurements," *International Journal of Distributed Sensor Networks*, vol. 2011, 2011.

[21] L. Lin, H.-C. So, F. K. Chan, Y. Chan, and K. Ho, "A new constrained weighted least squares algorithm for TDOA-based localization," *Signal Processing*, vol. 93, no. 11, pp. 2872–2878, 2013.

[22] H. Jamali-Rad and G. Leus, "Sparsity-aware multi-source TDOA localization," *Signal Processing, IEEE Transactions on*, vol. 61, no. 19, pp. 4874–4887, 2013.

[23] H. C. So, Y. T. Chan, and F. Chan, "Closed-form formulae for time-difference-of-arrival estimation," *Signal Processing, IEEE Transactions on*, vol. 56, no. 6, pp. 2614–2620, June 2008.

[24] A. Renaux, P. Forster, E. Boyer, and P. Larzabal, "Unconditional maximum likelihood performance at finite number of samples and high signal-to-noise ratio," *Signal Processing, IEEE Transactions on*, vol. 55, no. 5, pp. 2358–2364, 2007.

[25] R. Schmidt, "Least squares range difference location," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 32, no. 1, pp. 234–242, Jan 1996.

[26] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320 – 327, aug 1976.

[27] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proceedings of ICASSP 2008*, 31 2008-april 4 2008, pp. 2565 –2568.

[28] J. Velasco, C. J. Martín-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Processing*, vol. 119, pp. 209 – 228, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168415002650

[29] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[30] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1479–1489, 2008.

[31] P. Annibale, J. Filos, P. Naylor, R. Rabenstein *et al.*, "TDOA-based speed of sound estimation for air temperature and room geometry inference," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 234–246, 2013.

[32] M. Compagnoni, A. Canclini, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Source localization and denoising: a perspective from the TDOA space," *Multidimensional Systems and Signal Processing*, pp. 1–26, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11045-016-0400-9

[33] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3021–3028.

[34] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[35] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," *The quarterly journal of mathematics*, vol. 11, no. 1, pp. 50–59, 1960.

[36] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[37] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *in International Conference on Machine Learning*, 2011.

[38] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu, "Transduction with matrix completion: Three birds with one stone," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 757–765. [Online]. Available: http://papers.nips.cc/paper/3932-transduction-with-matrix-completion-three-birds-with-one-stone.pdf

[39] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theor.*, vol. 56, no. 5, pp. 2053–2080, May 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2044061

[40] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Procceedings of the MLMI*, ser. Lecture Notes in Computer Science, S. Bengio and H. Bourlard, Eds., vol. 3361. Springer-Verlag, 2004, pp. 182–195.

[41] D. C. Moore, "The IDIAP smart meeting room," IDIAP Research Institute, Switzerland, Tech. Rep., November 2004.

[42] G. Lathoud, "AV16.3 dataset," http://www.idiap.ch/dataset/av16-3/, [Last accessed in december 2015].

[43] X. Alameda-Pineda, "The gtde matlab toolbox." [Online]. Available: https://team.inria.fr/perception/the-gtde-matlab-toolbox/

[44] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

Jose Velasco received his Telecommunication Engineering Degree and Master Degree in Advanced Electronics, both from Universidad de Alcalá, Spain, in 2010 and 2011 respectively. Currently he is a Ph.D student and his research interests include audio and high dimensional signal coding and processing, with particular emphasis on sparse and low rank representation and modeling.

Daniel Pizarro received both his B.S. and PhD degrees from the University of Alcalá in 2003 and 2008 respectively. From 2005 to 2012 he was an Assistant Professor and member of the GEINTRA group at the University of Alcalá. From 2012 to 2014 he was an Associate Professor at Universite d'Auvergne and member of ALCoV group. Since 2015 he is Associate Professor at University of Alcalá. His research interests are in optimization, Computer Vision, including image registration and the reconstruction of deformable objects and their application to Minimally Invasive Surgery. Daniel Pizarro Pérez has co-authored more than 60 research papers; some in high-impact vision journals such as IEEE PAMI, IJCV, IMAVIS or CVIU and top ranking vision conferences, CVPR, ICCV or ECCV. He has participated in several research projects from public fundings from Spain and Europe and has leaded several technological transfer projects with Spanish companies.

Javier Macias-Guarasa received his Telecommunication Engineering Degree and PhD both from Universidad Politécnica de Madrid, Madrid, Spain, in 1992 and 2001 respectively. From 1990 to 2007 he was a member of the Speech Technology Group and held different teaching positions at Universidad Politécnica de Madrid, currently being Associate Professor in the Department of Electronics at Universidad de Alcalá, Madrid, Spain. In 2003 he was a visiting scientist at at the International Computer Science Institute (ICSI). He has authored or co-authored more than 100 referred papers in the field of Speech Technology and his current research interests are related to speech processing, audiovisual sensor fusion applications in intelligent spaces, and pattern recognition strategies applied to pipeline monitoring systems.

Afsaneh Asaei Afsaneh Asaei received the B.S. degree from Amirkabir University of Technology and the M.S. (honors) degree from Sharif University of Technology, in Electrical and Computer engineering, respectively. She held a research engineer position at Iran Telecommunication Research Center (ITRC) from 2002 to 2008. She then joined Idiap Research Institute in Martigny, Switzerland, and was a Marie Curie fellow on speech communication with adaptive learning training network. She received the Ph.D. degree in 2013 from École Polytechnique Fédérale de Lausanne. Her thesis focused on structured sparsity for multiparty reverberant speech processing, and its key idea was awarded the IEEE Spoken Language Processing Grant. Currently, she is a postdoctoral researcher at Idiap Research Institute. She has served as a guest editor of Speech Communication special issue on Advances in sparse modeling and low-rank modeling for speech processing and co-organized special issues on this topic at HSCMA'2014 and LVA/ICA'2015. Her research interests lie in the areas of signal processing, machine learning, statistics, acoustics, auditory scene analysis and cognition, and sparse signal recovery and acquisition.