Applied Acoustics 73 (2012) 698–712

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

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

Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments

J.M. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas, J. Escolano *

Multimedia & Multimodal Processing Research Group, Telecommunication Engineering Department, Polytechnic School, University of Jaén, Spain

ARTICLE INFO

Article history:
Received 8 July 2011
Received in revised form 26 November 2011
Accepted 6 February 2012
Available online 9 March 2012

Keywords:
Direction of arrival
Generalized cross-correlation
Spatially correlated signals

ABSTRACT

The localization of sound sources, and particularly speech, has a numerous number of applications to the industry. This has motivated a continuous effort in developing robust direction-of-arrival detection algorithms, in order to overcome the limitations imposed by real scenarios, such as multiple reflections and undesirable noise sources. Time difference of arrival-based methods, and particularly, generalized cross-correlation approaches have been widely investigated in acoustic signal processing, but there is considerable lack in the technical literature about their evaluation in real environments when only two microphones are used. In this work, four generalized cross-correlation methods for localization of speech sources with two microphones have been analyzed in different real scenarios with a stationary noise source. Furthermore, these scenarios have been acoustically characterized, in order to relate the behavior of these cross-correlation methods with the acoustic properties of noisy scenarios. The scope of this study is not only to assess the accuracy and reliability of a set of well-known localization algorithms, but also to determine how the different acoustic properties of the room under analysis have a determinant influence in the final results, by incorporating in the analysis additional factors to the reverberation time and signal-to-noise ratio. Results of this study have outlined the influence of the acoustic properties analysed in the performance of these methods.

1. Introduction

Localization of acoustic sources is a useful task in different scenarios and applications, such as separation of mixed audio signals, beamforming for suppressing noise of audio signals in a noisy environment, and pointing of cameras in video-conferences, underwater acoustics or human–machine interaction [1–9]. Although there is a general trend towards the use of a high number of microphones both in the research community and in the industry, the use of a pair of microphones is present in actual applications such as humanoid robotics [6,10,11], hearing aids [12] or modeling of psychophysical studies [13]. As pointed out by May et al. [12], because of the ability of the human auditory system, research dealing with binaural models of computational auditory scene analysis is a growing field. More recent works using just two microphones can be found also in [4,14,15]. The most challenging situations are those environments with moderate or high reverberation time and low signal-to-noise ratio.

Algorithms for acoustic localization can be divided into steered beamformers, high-resolution spectral estimation and time difference of arrival-based (TDOA) methods. TDOA-based methods are still dominant and they rely on relative delays between the two microphones [1]. Although TDOA based methods can be outperformed to a certain degree by more elaborate methods, they prove to have a great effectiveness due to their elegance and low computational costs [2]. Among them, Generalized Cross-Correlation (GCC) framework is one of the most successful approach of TDOA methods, since it was first introduced in [16]. This framework includes a wide range of algorithms for TDOA estimation. According to Refs. [3–5], they are described as indirect localization approaches, because they explicitly estimate the time delay of arrival before performing the localization task based on the knowledge of the microphones distribution. In GCC algorithms, the time delay between the signals in both microphones is estimated as the delay that, applied to one of the signals, maximizes the cross-correlation of both signals. The cross-correlation is usually weighted with a specific function. Thus, each of the GCC algorithms uses a different weighting function characterized by a particular behavior.

Several factors can affect the localization performance of the general cross-correlation methods: number of microphones, reverberation time, signal-to-noise ratio, number of sources, distance to the microphones, etc. The study presented herein is focused on the behavior of two microphones with a voice source in reverberant scenarios with stationary white noise. Despite the fact that a lot

* Corresponding author.
E-mail address: escolano@ujaen.es (J. Escolano).

0003-682X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2012.02.002

of research has been achieved with TDOA algorithms, not too much has been done in comparing the GCC family of methods when a pair of microphones is used in this kind of real environments. A summary of the analysed conditions and outcomes of previous evaluation studies is presented in Table 1. It is worthy to note that other existing techniques, such as microphone arrays with three or more transducers, have been evaluated in real rooms, whereas GCC algorithms with two microphones have been barely investigated under these situations. When the number of microphones is limited to a pair of them, the technical literature has compared the algorithms under highly unrealistic acoustic scenarios and empty rooms [27], simulated environments [28] or real scenarios without

Table 1
Evaluation conditions and main outcomes of research studies about performance assessment of GCC algorithms.

Study (Microphones)	Methods	Scenario	Conditions	Performance
Array usage				
Bartsch10 [17] (8/16)	(1) PHAT (2) AED[a]-PHAT (a) GCF[b] (b) LMS[c]	Laboratory	**Source:** five positions	Localization rate. The four methods (1a, 1b, 2a, 2b) can be used with an error inferior to 0.4 m
Brutti08 [18] (7)	(1) PHAT, (2) AED (a) GCF, (b) OGCF[d], (c) LS	Domestic like environment RT[e] = 0.65 s	**Source:** four speaker positions. **Noise:** BGN[f] and WN[g] (17, 7, 2 dB)	Localization rate (percentage of fine estimation) and accuracy (RMSE[h]). (1) is better than (2) for low SNRs. (c) is as accurate and precise as (a) and (b)
Cobos11 [5] (6)	(1) Proposed (2) SRP-HAT[i] (3) SRC[j]	(1) Simulated RT = 0.2, 0.7 s (2) Real scenario RT = 0.28 s	**Noise:** 0–10 dB. **Source:** 30 positions with different grid sizes	RMSE. Similar performance that SRP-HAT with less computational cost
Markovic10 [2] (4)	GCC-PHAT α + particle filtering	Class room RT = 0.6 s	**Source:** Speaker in movement	Detection reliability (accuracy >5°), RMSE, SD. Square array is better (97%, .4°, .3°)
Marti11 [19] (6)	Modified SRP-PHAT. Speech (S) and non S (NS) discrimination	Conference room. RT = 0.4 s	**Source:** 12 positions in two grid sizes	Percentage of well classified frames (S or NS) and rate of successful position estimation (>98%)
Mungamuru04 [20] (var-24)	(1) ML[k] (2) Weighted SRP-HAT (3) SRP-HAT	(A) Simulated (B) Laboratory (RT = 0.1 s)	**Source:** Moving speakers	Estimation accuracy. Average error and percentage of anomalies. ML-PHAT for position, ML for orientation
Omologo97 [21] (4)	PHAT-CM[l]	Real room RT = 0.3 s	**Source:** Whistle and speech in 15 positions. **Noise:** Coherent and BGN	Positions estimations with average location errors <10 cm
Ui-Hyun08 [22] (6)	GCC	Office room	**Noise:** Coherent[m] **Source:** θ [0–360]° with 10 steps. α three positions	Estimation Success rate (97.27%)
Valin03 [23] (8)	Modified GCC	Laboratory	**Source:** Speaker at four positions (distance, α)	Mean angular error. Precision of 3° over 3 m
Wang97 [24] (4)	GCC (onsets and BGN)-PictureTel	Conference room	**Source:** six positions	SD of positions (meters, Θ, α)
Yu04 [25] (16)	(1) LEMSag (2) LEMS improved	Real environments	SRE[n] [−2 dB, −12 dB], SNR [3 dB, 12 dB]	Micro separation analysis and optimal TDOA vectors size (N = 8). Better estimation success of (2)
Zhang08 [26] (6)	(1) SRP-PHAT (2) ML	Simulated RT = 0.1, 0.5 s	**Noise:** Coherent [0–25 dB]. **Source:** Speaker [0–360]° in 36° steps.	Percentage of estimations with accuracy >2° and >10°. Results better for (2) when SNR <25 dB
Two microphones				
Benesty00 [27]	(1) Proposed (eigenvalue) (2) PHAT (3) GCC (4) FC[o]	Varechoic chamber. RT = 0.15, 0.25, 0.74 s	**Source:** three positions (Speaker, WN)	Hits percentage. Different fail rates for the three positions and reverberations times
Murray04 [10]	GCC	Real scenarios	**Source:** speaker at 10 angles.	Average accuracy (<1.5°).
Rui04 [28]	Four weighted functions + 3 noise removal techniques	Simulated environment. RT = 0.05 s	**Noise:** Coherent. **Source:** Θ: 18 angles [10–170°]	Average error and SD. MLR + WG[p] and Wswitch + WG report better performance
Trifa07 [11]	(1) GCC (2) PHAT (3) MODD[q] (4) COCH[18]	Real scenario	**Noise:** Coherent (62.9 dB), music (76 dB), WN (79 dB). **Source:** Speaker at 60°	Average and SD of angular estimation. (2) has greatest accuracy and (3) is most reliable and precise
Yang09 [29]	(1) PHAT modified to remove reverb (2) PHAT	(1) Simulated. RT = 0–0.5 s (2) Conference room RT = 0.5 s	**Noise:** (1) WN SNR[0–40 dB] (2) Coherent (7 dB) **Source:** Speaker at 60°	RMSE results <10° for (1) with SNR <20 dB

[a] Adaptative Eigenvalue Decomposition (AED).
[b] Global coherence field SRP-PHAT (GCF).
[c] OGCF (Oriented GCF).
[d] LMS (Least Mean Squares).
[e] Reverberantion Time (RT).
[f] Background Noise (BGN).
[g] White noise (WN).
[h] Root Square Mean Error (RSME).
[i] Steering Response Power (SRP).
[j] Stocastic Region Constraction (SRC).
[k] Maximum Likehood (ML).
[l] Coherent measurement (CM).
[m] Coherent noise (Air conditioner and/or computers).
[n] Signal to Reberveration Energy (SRE); Maximum Likehood for Reverberation Wiener Filtering and Gnn substraction (MLR-WG).
[o] Fischell-Coker algorithm (FC).
[p] Moddenmeijer information theoretical approach (MODD)
[q] Cochlear filtering (COCH).

noise [10]. Some experiments in real scenarios with noise are presented in [11] and [29], but they are limited to a comparison in just one scenario. Also, the experiment in [29] is limited to one location of the voice sound with a time duration of 2.5 s and in [11] only one position of the voice is used as well.

Moreover, the technical literature usually characterize the scenarios acoustical properties of the scenarios with the reverberation time, without considering, for example, the nature of possible multiple reflections in the room, e.g. if they are specular or diffuse. Due to this lack of algorithms evaluation with real systematic acoustic measurements, the aim of this study is to compare the performance of different GCC algorithms with only one speech source, in different real environments characterized by specific room-acoustic parameters and with different *a priori* known signal-to-noise (SNR) ratios. As for dealing with noise sources, also many simulations have been performed considering unrealistic noisy situations by just adding uncorrelated noise directly to the microphone signals, which leads to totally uncorrelated and ideal noise. This kind of noise is typically due to internal/thermal noise of the measurement devices, but nowadays the use of medium/high quality acquisition instruments usually minimize this noise. In realistic situations, external noise sources such as those related to machinery, air conditioners, and electric motors are the most common ones. When being captured with two microphones, this kind of stationary noise is characterized by having a spatial correlation [28] and this is the kind of noise considered in the work presented here.

For these reasons, the main aim of this study is to evaluate GCC methods in typical conditions of real scenarios, that is to say, with the presence of external, stationary and in most cases, correlated noise sources. Furthermore, this study is based on the assumption that in order to evaluate accurately performance of these TDOA methods, in terms of the acoustic properties of a real scenario, it is necessary to consider other features together with the typical ones used, reverberation time and SNR.

The paper is organized as follows. Section 2 introduces a brief description of the evaluated methods. Next, the scenarios where these methods have been tested are described in Section 3. The methodology to perform the systematic acoustic measurements is presented in Section 4. Finally, the main contributions of this study are discussed and the conclusions outlined are presented in Sections 5 and 6, respectively.

2. GCC Framework

The original GCC framework is based on a single-path propagation model of acoustic plane waves emanating from a remote source and monitored at two separated microphones, where the signal acquired is a delayed and attenuated version of the original source with added noise [16,30]:

$$x_1(t) = \alpha_1 s(t+T) + n_1(t), \tag{1}$$

$$x_2(t) = \alpha_2 s(t+T+\tau) + n_2(t), \tag{2}$$

being $x_1(t)$ and $x_2(t)$ the signals at both microphones, $s_1(t)$ the original acoustic signal, α_1 and α_2 the attenuation factors due to propagation, $n_1(t)$ and $n_2(t)$ the noise at both microphones. The time T is the delay of the path between the acoustic source and the first microphone, while τ is the *time difference of arrival*.

The propagation model can be described in a more general way, if the room impulse responses at the locations of the microphones are considered [3]:

$$x_1(t) = a_1(t) * s(t) + n_1(t), \tag{3}$$

$$x_2(t) = a_2(t) * s(t) + n_2(t), \tag{4}$$

being $*$ the time convolution operator, $a_1(t)$ and $a_2(t)$ are the room impulse responses at their corresponding source/receiver positions.

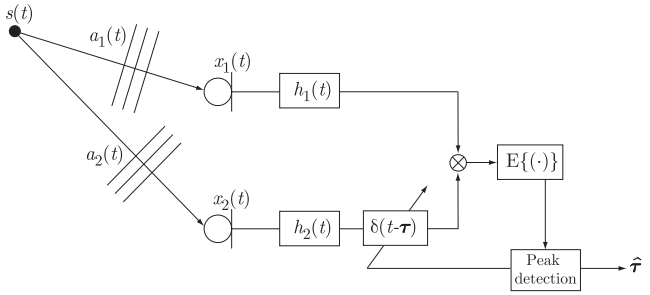

Fig. 1. Block diagram of a generalized cross-correlator for TDOA estimation [3].

In this model, the TDOA is implicitly found in the difference of these responses; and the optimal time delay can be obtained with a generalized cross-correlator (Fig. 1) as

$$\hat{\tau} = \arg\max_{\tau} E\{(x_1(t) * h_1(t))(x_2(t+\tau) * h_2(t))\}, \tag{5}$$

being $E\{(\cdot)\}$ the statistical average over time. Using this correlator, the optimal delay $\hat{\tau}$ converges to the time difference of arrival between signals $x_1(t)$ and $x_2(t)$. Impulse responses $h_1(t)$ and $h_2(t)$ are the weighting functions applied to signals $x_1(t)$ and $x_2(t)$. GGC methods use particular weighting functions with different behavior, and therefore, their estimation performance may be different in the same scenario.

If the *generalized cross-correlation function* is defined as

$$\varphi^g_{x_1x_2}(\tau) = E\{(x_1(t) * h_1(t))(x_2(t+\tau) * h_2(t))\}, \tag{6}$$

then Eq. 5 can be rewritten as:

$$\hat{\tau} = \arg\max_{\tau} \left\{\varphi^g_{x_1x_2}(\tau)\right\}. \tag{7}$$

Also, if Eq. 6 is written in the frequency domain, the generalized cross-correlation can be related to the cross power spectral density function $\Phi_{x_1x_2}$ [16]:

$$\varphi^g_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} H_1(f)H_2^*(f)\Phi_{x_1x_2}(f)e^{j2\pi f\tau}df \tag{8}$$

To further simplify the GCC framework, the term *generalized frequency weighting* [16] is commonly defined as:

$$\psi_g(f) = H_1(f)H_2^*(f). \tag{9}$$

Thus, the methods included within the GCC framework can be defined by just specifying the generalized frequency weighting $\psi_g(f)$ used. In this work, the methods CC (*Cross-Correlation*) [31], PHAT (*Phase Transform*) [32], Roth (*Wiener–Hopf weighting*) [33] and HT (*Hannan-Thomson Maximum Likelihood*) [34] have been compared. Their respective weighting functions are listed in Table 2 [3,16].

In the case of not using any weighting function, the method is just known as *cross-correlation*, and it is the most straightforward algorithm for estimating the delay between two signals. The other

Table 2
Weighting functions of the tested methods. The term $\Gamma(f)$ represents the *coherence* between the microphone signals.

Approach	Weighting function $\psi_g(f)$
Cross-Correlation (CC)	1
Phase Transform (PHAT)	$\frac{1}{\lvert\Phi_{x_1x_2}(f)\rvert}$
Roth (Wiener–Hopf weighting)	$\frac{1}{\Phi_{x_2x_2}(f)}$
Hannan–Thomson (maximum likelihood estimate)	$\frac{1}{\Phi_{x_1x_2}(f)} \cdot \frac{\lvert\Gamma(f)\rvert^2}{1-\lvert\Gamma(f)\rvert^2}$

methods of the GCC framework differ in the specified function. In Roth method, the weighting function allows suppressing the frequencies where the power spectrum of the additive noise is large, and therefore, where the estimation of the cross-correlation may be erroneous [16]. This can lead to a more accurate delay indication than in the case of using just the CC method. PHAT is popular for having a good behavior in reverberant environments with low noise. It uses the magnitude of the cross-power spectral density of both signals as the weighting function and, despite of being developed as a heuristic approach, it has been shown to be robust under reverberation in low noise environments [26,29,35]. Also, it is theoretically proved that this method eliminates a spreading effect that occurs due to the existence of an uncorrelated noise at both microphones [16]. The HT method estimates an optimal delay from a statistical point of view under conditions of an ideal acoustic propagation, since the estimation variance can achieve the Cramer–Rao lower bound in those conditions. In this case, the weighting function is based on the *coherence function* $\Gamma(f)$ between signals $x_1(t)$ and $x_2(t)$, which is defined as [3]:

$$\Gamma(f) = \frac{\Phi_{x_1x_2}(f)}{\sqrt{\Phi_{x_1x_1}(f)\Phi_{x_2x_2}(f)}} \tag{10}$$

3. Scenarios description

This study is based on the assumption that along with the reverberation time other architectural acoustic parameters related to the wall absorbing/scattering properties must be considered in order to evaluate estimation performance of TDOA methods. This is hypothesized based on the diffuse sound field theory [36], where under the assumption of perfectly diffuse boundaries, the cross-correlation between pressure measurements at two different points is very low whether the distance exceeds half of the wavelength [37] and the sound field reaches stationarity. It is well-known in architectural acoustics, that a scenario with these features is the only physical way to obtain nearly spatial uncorrelation between two near measuring points [38]. However, as previously described, TDOA methods performance is conditioned by the uncorrelation property of the noise registered in the microphones.

A set of scenarios has been selected to assess the extent to which their absorbing/scattering properties have a clear influence in the reliability of these algorithms, specially attending to those aspects that may alter the uncorrelated noise assumption, since perfectly diffuse surfaces are physically unachievable. In particular, the estimation performance of TDOA methods has been tested in the next scenarios: a lecture room with high reverberance and moderate specular reflections, an office with a variable reverberation time and a considerable amount of scattering objects, and an auditorium furnished with absorbing materials. The particular acoustic properties of each scenario are described in detail in the subsequent sections. The parameters of the scenarios related to reverberation time, early decay time, nature of the first reflections and absorption have been listed in Table 3. They are volume V, total surface S, broadband reverberation time RT measured according to [39] (see Section 4.3 for details), early decay time EDT, ratio early decay time to reverberation time EDT/RT and averaged absorption coefficient $\bar{\alpha}$, estimated by applying Norris–Eyring formulae [40] with the measured reverberation time.

Table 3
Acoustic properties of the analyzed scenarios. The ratio Early Decay Time to the Reverberation Time (EDT/RT) indicates if the first reflections are diffusive (ratio close to 100%) or specular. Based also on the measured RT, the averaged absorption coefficient $\bar{\alpha}$ is estimated using the Norris–Eyring formulae.

	V (m^3)	S (m^2)	RT (s)	EDT (s)	EDT/RT (%)	$\bar{\alpha}$
Lecture room	473	418.2	1.91	1.79	93.89	0.091
Office	118	149.8	0.46	0.47	102.16	0.241
Modified office	118	149.8	0.39	0.39	100	0.278
Auditorium	740	626	0.54	0.47	86.42	0.297

Classically, GCC-based methods performance has been evaluated in terms of the reverberation time of the scenario where the measurements are carried out. The aim of using this parameter is, somehow, to provide a measurement of how the probability of the estimated direction of arrival varies with the amount of strong reflections. However, this parameter poorly provides information about how first reflections are, either – mostly – specular or diffuse. For this reason, not only the reverberation time parameter is evaluated, but also the ratio EDT/RT [41]. This parameter can be seen as a measurement of the directness, being its typical values between 0.8 and 1.1. If surrounding surfaces direct early reflections onto measurement points, this reduces the early decay time, giving a low ratio, which can be interpreted as first reflections being mostly specular. However, the closer this value is to unity, the more diffusive are considered these first reflections.

3.1. Lecture room

The first scenario corresponds to a typical lecture room. It has a volume of 473 m^3 and a total surface of 418.2 m^2. The wall surfaces are entirely made with plaster, gypsum and glass. Thus, these surfaces could be considered as even in the entire speech frequency band. Ceiling and walls are made of plaster and floor is completely covered with marble. This room is full of tables and chairs, all made of plastic, being the only scattering objects of the room. Thus, most of the room elements are expected to produce specular reflections. Indeed, the analysis performed of this room (see Table 3) has identified it as a highly reverberant scenario with a considerable amount of first specular reflections and poor absorption. The parameter EDT/RT indicates that the nature of the first reflections in this scenario is more specular than diffusive.

3.2. Office

The office is a fully fitted room with a volume of 118 m^3 and a total surface of 149.8 m^2. The furnishing elements cover at least one third of the wall surfaces, and there are several tables and chairs that cover the floor. These furnitures are filled with a considerable amount of office elements, which make that most of the reflecting surfaces may be considered as diffusively reflecting surfaces. The visible wall surfaces and ceiling are made of plaster, whereas the floor is covered with marble. This scenario can be described as a room with low absorption, a considerable amount of scattering surfaces and moderate reverberance (see Table 3). The parameter EDT/RT confirms that the first reflections in this scenario are of diffusive nature. Thus, it is an excellent scenario to analyse the effect of the expected partially uncorrelated noise recorded at the two microphones.

3.3. Modified office

As previously mentioned, the existence of a high number of scatterer objects in the office (described on Section 3.2) makes it an ideal scenario for analysing the effect of partially uncorrelated noise. However, it would be also interesting to reduce its reverberation time, in order to make measurements in a highly diffusive room but with medium–low reverberation time (see Table 3). With this purpose, the previous office was modified by covering floor and plaster surfaces with carpets and synthetic materials, while the scatterer objects (furniture and books) were kept the same. This new scenario is referred from now on as Modified Office. As can be seen in Table 3, the absorption coefficient has been

increased while maintaining the scattering reflections, as it is indicated by the ratio EDT/RT.

3.4. Auditorium

The fourth scenario is an auditorium, with 740 m^3 of volume and a total surface of 626 m^2. The main characteristic of this room is the considerable amount of absorbing material in a medium size auditorium. Despite of being full of scattering surfaces, such as chairs and decorative elements, most of them were covered with heavy cotton cloths and curtains, and the chairs were uphostered. Thus, the reverberance of this scenario is moderate despite of its high volume (see Table 3). This scenario is of special interest because it presents a similar reverberation time than the office, but the first reflections are specular as indicated by the parameter EDT/RT, while they are difussive in the office. If the methods were evaluated taking into account only the reverberation time and the SNR, the auditorium and the office would present the same behavior. However, in Section 5 it will be shown that the influence of the noise is different in both scenarios.

4. Method

The experimental setup is described in this section, detailing the measurement procedure followed and the data analysis.

4.1. Experimental setup

Fig. 2 depicts the main elements of the experimental setup: two Behringer B-5 omnidirectional microphones connected to a PC via a firewire port with a MOTU-8pre audio interface, a multimedia speaker Wunderton CS-6501 placed at a known angle θ_S respect to the pair of microphones for the voice source and an omnidirectional speaker AVM DO-12 placed at a fixed location θ_N for the noise source. Power levels of voice and noise sources were controlled by two power amplifiers InterM CM-10.5. The experiment has been done with four levels of SNR ratio: three levels with 10 dB, 20 dB and 30 dB using the noise source via measuring and controlling the sound pressure level at each microphone with a sound level meter Rion NL-32 (configured in slow mode and no filtering), and a fourth level with the ambient noise of each scenario. In this last case, the noise varies among the scenarios because it is not controllable, being the SNR in a range among 40–45 dB.

The generalized cross-correlation methods have been tested in the scenarios described in Section 3 and with different directions of arrival: 30°, 60° and 90°. This last angle corresponds with the situation where the loudspeaker is placed in the normal direction to the plane containing the two microphones. The microphones were placed at the center of each room. For the sound/voice source, an anechoic male voice available at the *First Stereo Audio Source Separation Evaluation Campaign* [42] has been used. The voice was reproduced three times per recording, to have three repetitions of each experiment. The repetitions were separated by a convenient pause to avoid undesirable mixing effects due to reverberance. Thus, results were analyzed averaging the values obtained from each repetition.

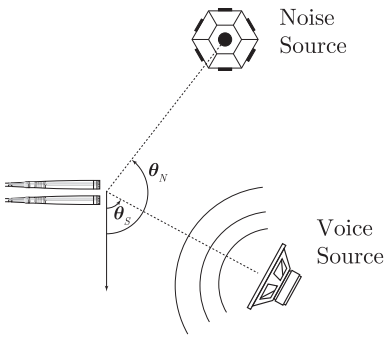

Fig. 2. Experimental setup for the scenarios. The voice source is located at different angles respect to the microphones, while the noise source is fixed.

In order to produce additive gaussian white noise, a pseudo-random sequence low-pass filtered with 5 kHz was used as the noise source. This noise was also emitted without any other signal during 3 s and before each of the voice signal repetitions, in order to assure that it reaches its stationary stage in the room.

4.2. Algorithms implementation

Fig. 3 depicts the implementation of the tested methods. The audio signals were captured at the microphones with a sampling frequency of 96 kHz and low-pass filtered with a cutoff frequency of 5 kHz. Thus, the signals were processed just within the speech frequency band. The direction of arrival was estimated for analysis frames of 25 ms, windowed with a Hann window. The cross-power spectral-density function was obtained in each window using 512 samples Fourier transforms via FFT, and the cross-correlation was computed using its corresponding inverse Fourier transform via IFFT. An interpolation stage using natural cubic splines has been done to search the maximum peak in the correlation function. Thus, a better angular resolution is achieved. For each analysis window, a TDOA value was obtained and the corresponding angle estimated as (see Fig. 4):

$$\cos\theta_S = \frac{c \cdot \hat{\tau}}{d} \tag{11}$$

being c = 343 m/s the speed of sound and d the separation between microphones. The microphones separation was 7.9 cm. If the microphone separation is increased, the localization performance may improve due to a better angular resolution and because the correlation of the noise is decreased. However, for large microphone separations, the peak in the cross correlation of the voice signals is more diffuse, which makes its detection less reliable [25]. On the other hand, the purpose of the experiments was to compare the methods under a common arrangement, more than optimizing this arrangement. Respect to the possible spatial aliasing, as pointed out in [30], it has to be treated with great care in the context of beamforming and noise reduction, but is not a big concern for the task of source localization.

Performance of the algorithms was assessed in 10 s intervals of the audio signal recorded. Fig. 5a depicts with dots the estimated positions of the audio source for each window of 25 ms. The normalized signal energy has been superimposed in the same figure. This energy curve allows to appreciate the different levels of the captured sound in time domain and its peaks highlight those windows where the voice signal is predominant. Therefore, with the purpose of selecting the frames with a certain level of voice energy, the estimations obtained in windows with an energy level under a previously established threshold were discarded. This is shown in Fig. 5b, where only those positions with an energy level over a threshold are drawn. The threshold level used in the experiments corresponds to the average energy level of the overall signal acquired (dashed line in Fig. 5b).

For a better visualization of the results, a histogram which represents probability function $f_{\Theta}(\theta)$ for the source localization has been computed for each experiment. These histograms can be found at the end of the paper as an Appendix A. The histogram is a graphical representation of the probability function estimation of a variable, which is built by accumulating its measured values. In this particular case, the histogram is built upon the number of

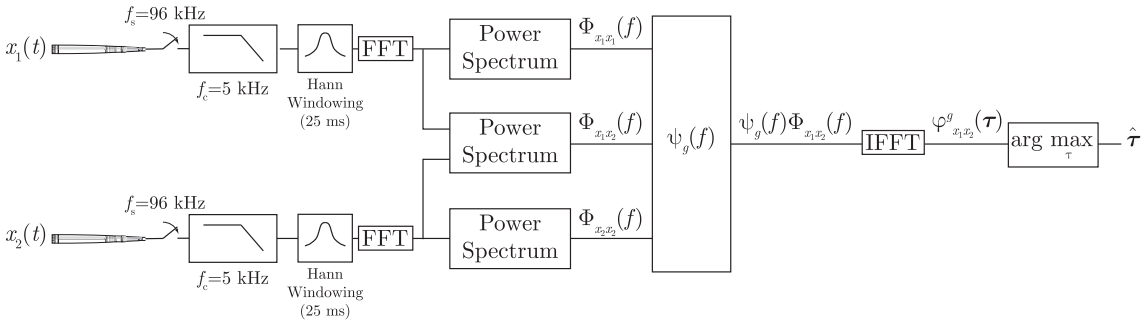

Fig. 3. Block diagram for the tested methods.

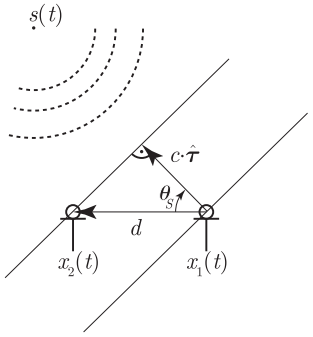

Fig. 4. Angle of incidence of a plane wave and $\hat{\tau}$. Angle θ is the speaker location.

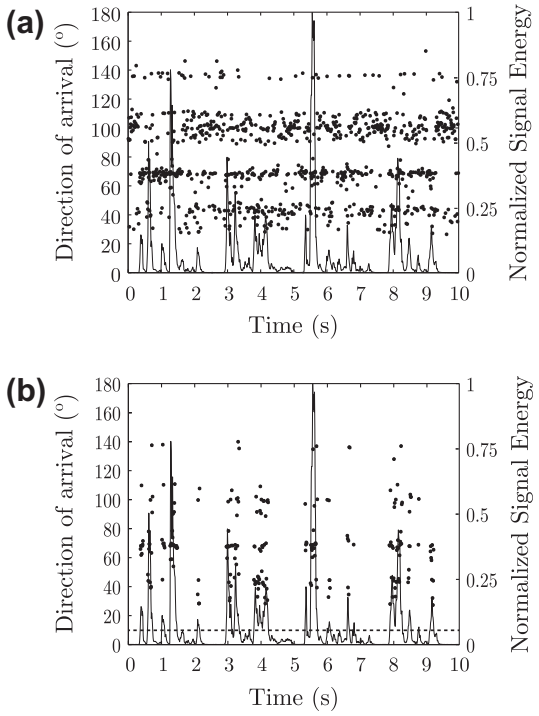

Fig. 5. (a) Estimated angles (dots) and normalized signal energy (line) for an analysis time of 10 s with the PHAT method in the office scenario. The sound is located at 60° and the SNR is 20 dB. (b) In this case the dots represent only those angles that will be accumulated in the histogram.

times an specific angle is estimated. The shape of the histogram sometimes is particularly sensitive to the number of bins. If the bins are too wide, important information might get omitted. For example, the data may be bimodal but this characteristic may not be evident if the bins are too wide. On the other hand, if the bins are too narrow, what may appear to be meaningful information really may be due to random variations that show up because of the small number of data points in a bin. The number of bins has been set as the square root of the number of observations [43], in order to obtain smooth histograms without losing relevant information. For the experiments presented here, the number of bins has been among 30 and 32 for each histogram.

For obtaining a more accurate estimation, angles with high energy level were given a greater weight in the accumulating process. Therefore, the mode – i.e. the maximum value in the histogram – is located at the most probable angle and, applied to a voice source localization scenario, it is used to estimate the most likely source position [3,4,30]. Besides the mode, three more statistics are computed in order to analyze the performance of the algorithms in a more quantitative way: the relative frequency of the mode, which informs what fraction of the time the mode is detected; the mean, which can be useful in order to know if the distribution is unimodal when it is compared to the mode; and the root mean square error (RMSE), which indicates the overall accuracy of the algorithm respect to the real direction of the voice source. The performance of the methods will be analyzed taking into account these four statistics in Section 5, as they can give clearer information than the histograms.

In order to characterize the properties of the noise, several intervals have been recorded with only the noise source emitting in the scenario, and the normalized cross-correlation between the signals has been computed. In order to determine the degree of spatial correlation between both noise signals, the maximum value of this normalized cross-correlation $\max\{\bar{\varphi}_{n_1n_2}(\tau)\}$ is used as an estimator. A more detailed description of this analysis is further presented at Section 5.

4.3. Decay curve measurement

Each scenario has been characterized through the impulse response measured with the pair of microphones at the same location where the experimental recordings were made. Two methods were used to obtain the room impulse response: Maximum Length Sequences [44] and Swept-Sine techniques [45]. The sound was emitted with the omnidirectional speaker AVM DO-12 and previously amplified by one of the power amplifiers. Then the decay curve was computed using just one of the two impulse responses by applying a backward Schroeder's integration [39,46]. Furthermore, based on the obtained decay curve, the broadband reverberation time and early decay time parameters were computed according to [39]. These parameters, along with the volume and total surface of the rooms, were used as explained in Section 3 in order to characterize acoustically the scenarios.

5. Results and discussion

This section presents the systematic acoustic measurements performed in the four different scenarios described in Section 3, and the results after processing them. In the following subsections, results obtained in each scenario are analyzed in detail. A final subsection summarizes the results and conclusions obtained through the different analysis.

5.1. Lecture room

The two microphones are placed at the middle of the lecture room, whereas the source loudspeaker is located at a constant distance of 4 m approximately. The noise source is placed in an angle of 145° with respect to the microphones.

The statistics from the data obtained in the experiments are shown in Table 4, where the four methods are compared. Each subtable represents a single situation, where the sound source is placed at a specific angle for a given SNR in the lecture scenario. The mode and mean values of an ideal histogram should be the angle where the sound source is placed, whereas the frequency (of the mode) should be near to 1, and the error near to 0. In general, all the methods were affected by the multiple reflections of the environment and the presence of noise.

In this kind of scenario, two negative side effects are expected. First, reverberation time is strongly related to the amount of reflections and how they remain in the room during a certain time, even when the sound source is deceased. This effect makes clearly lower the relative frequency of the mode, and the RMSE value increases. Second, at the level of SNR where the noise disturbs the localization estimation of the desired source, the RMSE also increases due to the appearance of observations corresponding to the direction of the noise source, the mean moves away from the mode indicating that the distribution is not unimodal, and the relative frequency also gets lower. In the worst case, the value of the mode becomes closer to the position of the noise source, and the mean is nearer the noise position than the voice one.

The *lecture room* scenario is characterized by a high reverberation time (approximately 2 s), and it is a good example to see how multiple reflections affect results obtained when there is just ambient noise in the room. When the voice source is located in front of the two microphones, i.e. the angle is 90°, the four methods have a good performance. However, when the source separates from 90°, clearly the CC method is the most negatively affected, being the value of the mode far from the real direction of arrival, and its relative frequency low, indicating that the distribution of the observations is spread over a wide range of values. This can be explained by the fact that the weighting functions used in the other methods undermine the effects of low frequencies, which are typically more affected by the reflections than the higher ones. This effect leads to an improvement in the time delay estimation [3]. From the architectural acoustic point of view, this effect seems to be reasonable since most materials have a considerable absorption coefficient at medium–high frequencies whereas at lower frequencies, materials are less absorptive.

When comparing the detection performance for the three source positions under the same conditions, the mean square error indicates that the performance of the estimation decreases when the angle separates from 90°, independently of the SNR level. This effect is also pointed out in Ref. [28], although through simulations performed by using the image method [47]. In general, all the methods have their best performance when the sound source is placed at 90°, that is to say, when the audio signals in both microphones are nearly identical. In positions with an azimuth angle close to 0° or 180°, the estimation performance is lower, not only limited by the frequency sampling and the microphones distance, but also due to the non-lineal transformation of Eq. 11 [2,22].

Respect to the influence of the noise, the statistics show how the performance clearly gets worse for PHAT, HT and Roth methods with the decrease of SNR. Thus, the frequency of the mode decreases, the mean separates from the mode, and the RMSE increases. For a SNR of 10 dB, when the voice source is far from 90°, the mode indicates that the direction of the noise source (143.4°) predominates in the estimations (although the mean shows that the distribution is not unimodal, as the voice source is present in the observations). On the other hand, it is worthy to note that the CC method was not so clearly affected, although it shows a very low performance in this scenario (see mode values at Table 4), mainly due to the high reverberation. Attending to the mode and its relative frequency, it can be concluded that in this scenario conditions PHAT outperformed the other methods with a SNR above 10 dB. However, RMSE was high due to the influence in the estimations of the noise source, which was placed far from the voice source, at 145°. As the noise source was emitting during all the experiment duration and the voice source had silences, the contribution of noise was higher for low conditions of SNR. For this reason, the value of the mode was computed as the position of the noise source with a SNR of 10 dB.

Table 4
Lecture room statistics. The four methods are compared for each combination of the sound source position and SNR. *Mode* is the most probable angle according to the method, *Frequency* is the relative frequency of the mode, *Mean* is the mean value of the observations, and *RMSE* is the root mean square error respect to the real direction of arrival. In this scenario, the noise source is located at 145°. The best method has been highlighted for each situation attending to the mode value and its relative frequency.

Approach	Ambient noise				SNR = 30 dB				SNR = 20 dB				SNR = 10 dB			
	Mode	Fequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE
Angle = 90°																
PHAT	**87.2**	**0.53**	**89.5**	**5.5**	87.2	0.46	89.3	11.6	87.2	0.36	91.3	19.8	92.8	0.15	101.4	31.1
HT	87.2	0.53	89.5	8.2	92.8	0.40	90.0	15.7	87.2	0.31	91.5	21.3	92.8	0.11	94.7	29.9
Roth	87.2	0.51	88.7	10.9	87.2	0.45	88.6	14.1	87.2	0.36	91.6	21.5	104.1	0.15	96.8	30.1
CC	87.2	0.51	89.6	5.6	**87.2**	**0.51**	**89.6**	**5.4**	**87.2**	**0.47**	**89.8**	**5.9**	**92.8**	**0.35**	**92.1**	**8.2**
Angle = 60°																
PHAT	**59.1**	**0.84**	**62.4**	**13.9**	**59.1**	**0.58**	**65.7**	**18.3**	**59.1**	**0.45**	**78.7**	**36.7**	143.4	0.22	103.9	60.3
HT	59.1	0.76	62.6	14.8	59.1	0.47	66.6	21.2	59.1	0.37	70.4	29.0	143.4	0.17	97.4	54.0
Roth	59.1	0.80	62.3	14.9	59.1	0.53	65.2	18.1	59.1	0.39	76.0	36.3	**59.1**	**0.19**	**86.4**	**47.4**
CC	70.3	0.25	75.7	19.2	70.3	0.24	75.9	19.1	75.9	0.27	76.3	19.7	70.3	0.27	76.6	20.6
Angle = 30°																
PHAT	**30.9**	**0.40**	**49.3**	**36.6**	**30.9**	**0.24**	**60.9**	**50.4**	**30.9**	**0.23**	**77.5**	**65.7**	143.4	0.26	106.2	88.2
HT	30.9	0.33	53.7	43.6	30.9	0.17	63.3	51.3	30.9	0.15	79.9	68.0	143.4	0.18	97.5	80.1
Roth	30.9	0.31	53.1	42.5	30.9	0.18	63.9	50.3	30.9	0.12	78.4	66.3	143.4	0.14	95.8	76.9
CC	81.6	0.16	75.1	48.5	87.2	0.14	74.5	48.0	87.2	0.15	75.0	48.6	**87.2**	**0.15**	**76.6**	**50.4**

Regarding to the noise *quality*, the algorithms analysed assume uncorrelated noise recorded at each microphone. Ideally, this can be only achieved within an ideal completely diffuse sound field, where the energy density is uniformly distributed in space and where the energy flow is isotropic. In these conditions, the constituent plane waves are uncorrelated [38]. This unrealistic scenario might only be achieved if all the room surfaces at the room are totally diffuse. In this first scenario, the lecture room, a strong influence of the diverse reflections is expected, due to the low-absorptive materials that characterize this room and moderate volume (see description at Section 5.1). Attending to the walls, ceiling and floor, it could be assumed that most of the reflections were specular since they were built with even surfaces. However, the high number of furnished elements -tables and chairs- and the high reverberance in the room turn the reflections to be diffused after certain time [40], making noise becomes more uncorrelated.

The uncorrelation of the noise was evaluated by analyzing those intervals where only noise was present. Taking samples for 1.5 s., the maximum of the normalized cross-correlation of the noise signals in the two microphones was computed, indicating a moderate correlation, with $\max\{\bar{\varphi}_{n_1 n_2}(\tau)\} = 0.55$, between the noise captured in each microphone. Therefore, the realistic noise at these experiments differs from those where just an uncorrelated white noise at both microphones is directly added [3,30] to the measured voice signals at each microphone. Experiments where noise sources have been placed externally in a simulated or real environment are closer to this situation. However, those experiments are usually carried on in empty or simulated rooms and they usually take into account only the reverberation time of the scenario. Thus, the effect of the diffusion degree of the reflections has not been evaluated, meanwhile our study shows that it has considerable consequences in the final performance of the methods.

5.2. Office

Table 5 lists results of the office scenario, where the same procedure described in Section 5.1 has been followed. In this scenario, the noise source was placed at 135° with respect to the two microphones. This room is characterized by a moderate reverberance, low absorption and a considerable amount of scattering surfaces, as described at Section 3.2. According to the statistics, in ambient noise conditions, all the methods except CC showed an accurate estimation. Method CC estimated a proper arrival direction for the case of 90°, which is the position that can be more accurately estimated. However, the estimation in terms of the mode and its relative frequency was mistaken for 60° and 30°, specially in this last case, where the mode is far from the expected position, and its relative frequency is much lower than the rest. Again, this effect is explained because CC method has got more difficulties when working in reverberant environments, as previously pointed out.

Results when the noise source was introduced lead to worse estimations, the relative frequency of the mode became lower and the RMSE clearly increased for PHAT, HT and Roth methods when the SNR decreased. Again, in this scenario, the CC method was less affected by the noise than the rest of the methods, although its estimation were worst when the voice source was located at 30°. Among the other methods, the main differences in performance arised for the source location at 30°. For this condition, the PHAT estimation obtained better results attending to mode frequency and RMSE. However, for a SNR of 10 dB, the mean in the three methods differed from the mode value, indicating that the distribution was not unimodal, clearly due to the high presence of noise. With this low value of SNR (10 dB), the estimation of ROTH was closer to the true positions of the voice source for 60° and 90° showing that this method seems to be less sensitive to this noise.

Regarding the performance differences between scenarios, RMSE was in general higher for the lecture room than in the office, due to the negative influence of a higher reverberation time, affecting the multiple reflections the overall performance.

The analysis of the noise introduced in the scenario shows also the existence of a moderate correlation, with $\max\{\bar{\varphi}_{n_1 n_2}(\tau)\} = 0.52$. The maximum of the normalized cross-correlation value is not significantly different from the one in the lecture room scenario. This can be explained because, despite the presence of the high amount of scattering surfaces (which could lead to consider significantly uncorrelated noise signals), the ceil and floor, together with some naked wall portions, constitute a strong presence of specular surfaces, and the room is far to be considered totally diffuse. Thus, there exists a remarkable correlation for this medium reverberant scenario, similar to the previous one. However, this scenario differs considerably from the lecture room regarding to the boundary reflective characteristics. Whereas the lecture room has considerable amount of strong – and very located – first reflections, the office scenario provides a higher number of diffuse reflections contributing to uncorrelate the noise signals at microphones despite of having a smaller volume than the lecture room.

Table 5
Office statistics. The four methods are compared for each combination of the sound source position and SNR. *Mode* is the most probable angle according to the method, *Frequency* is the relative frequency of the mode, *Mean* is the mean value of the observations, and *RMSE* is the root mean square error respect to the real direction of arrival. In this scenario, the noise source is located at 135°. The best method has been highlighted for each situation attending to the mode value and its relative frequency.

Approach	Ambient noise				SNR = 30 dB				SNR = 20 dB				SNR = 10 dB			
	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE
Angle = 90°																
PHAT	87.0	0.55	88.7	4.9	92.8	0.52	93.9	11.3	99.0	0.32	91.8	9.7	105.0	0.32	104.5	24.0
HT	**87.0**	**0.58**	**88.5**	**7.0**	92.8	0.43	96.3	16.1	93.0	0.36	84.8	22.3	117.0	0.31	107.4	27.2
Roth	87.0	0.54	90.8	10.0	92.8	0.53	92.4	14.8	99.0	0.30	93.8	15.0	99.0	0.36	96.9	24.5
CC	87.0	0.57	89.7	5.5	**87.2**	**0.54**	**89.6**	**5.4**	**87.0**	**0.47**	**89.1**	**5.5**	**87.0**	**0.39**	**89.6**	**6.0**
Angle = 60°																
PHAT	63.0	0.48	61.8	11.8	66.8	0.26	67.9	18.7	66.8	0.25	71.6	25.6	99.0	0.24	71.4	29.5
HT	63.0	0.50	61.3	14.7	**61.0**	**0.28**	**66.5**	**18.4**	66.8	0.27	67.5	25.4	39.0	0.20	75.1	38.5
Roth	**63.0**	**0.59**	**67.7**	**21.9**	66.8	0.26	68.5	17.8	**61.0**	**0.28**	**74.6**	**26.9**	69.0	0.19	75.1	32.1
CC	69.0	0.36	73.2	16.6	72.6	0.27	71.5	15.8	61.0	0.22	73.6	17.5	**69.0**	**0.39**	**73.1**	**17.0**
Angle = 30°																
PHAT	**31.9**	**0.57**	**32.6**	**14.7**	**31.9**	**0.51**	**36.3**	**19.5**	**33.0**	**0.42**	**44.2**	**27.8**	**31.9**	**0.32**	**64.8**	**55.0**
HT	31.9	0.49	32.8	17.1	31.9	0.41	37.4	22.4	33.0	0.29	49.6	34.5	31.9	0.31	65.6	56.2
Roth	31.9	0.52	34.5	19.4	31.9	0.46	40.4	26.6	33.0	0.38	48.0	33.5	31.9	0.25	68.6	56.8
CC	55.2	0.21	58.1	31.0	61.0	0.27	58.6	31.5	57.0	0.28	58.5	31.2	66.8	0.23	60.1	33.1

Table 6
Modified office statistics. The four methods are compared for each combination of the position of the sound source and SNR. *Mode* is the most probable angle according to the method, *Frequency* is the relative frequency of the mode, *Mean* is the mean value of the observations, and *Error* is the root mean square error respect to the real direction of arrival. In this scenario, the noise source is located at 135°. The best method has been highlighted for each situation attending to the mode value and its relative frequency.

Approach	Ambient noise				SNR = 30 dB				SNR = 20 dB				SNR = 10 dB			
	Mode	Frequency	Mean	Error	Mode	Frequency	Mean	Error	Mode	Frequency	Mean	Error	Mode	Frequency	Mean	Error
Angle = 90°																
PHAT	**92.8**	**0.82**	**91.8**	**5.2**	**92.8**	**0.67**	**91.8**	**10.5**	98.4	0.62	96.2	13.3	113.2	0.50	104.8	17.5
HT	92.8	0.79	91.0	7.3	92.8	0.59	91.2	12.8	98.4	0.47	93.1	18.1	113.2	0.55	104.7	19.3
Roth	92.8	0.81	91.7	5.1	92.8	0.66	90.8	14.7	98.4	0.46	94.2	18.5	95.8	0.50	104.0	19.6
CC	87.2	0.46	88.3	5.1	87.2	0.47	88.5	5.0	**87.2**	**0.51**	**89.0**	**5.3**	**90.0**	**0.38**	**89.5**	**7.2**
Angle = 60°																
PHAT	**61.0**	**0.77**	**63.3**	**9.0**	**63.0**	**0.58**	**69.7**	**23.6**	69.0	0.23	72.7	25.0	101.6	0.36	81.9	33.6
HT	61.0	0.76	64.0	10.9	63.0	0.55	67.1	17.4	69.0	0.28	70.7	26.0	101.6	0.21	81.5	37.2
Roth	61.0	0.76	63.6	10.3	63.0	0.54	68.1	19.9	69.0	0.17	71.2	27.4	66.8	0.15	71.8	29.5
CC	66.8	0.32	74.9	17.4	69.0	0.37	74.9	17.5	**69.0**	**0.38**	**74.4**	**16.8**	**66.8**	**0.32**	**76.3**	**20.0**
Angle = 30°																
PHAT	**30.9**	**0.85**	**31.8**	**10.3**	**30.9**	**0.51**	**36.5**	**20.9**	31.9	0.29	50.2	37.5	**37.7**	**0.36**	**59.3**	**43.1**
HT	30.9	0.76	32.2	11.9	30.9	0.42	39.5	26.7	31.9	0.27	54.7	43.8	37.7	0.23	61.3	46.8
Roth	30.9	0.74	31.9	11.4	30.9	0.49	37.2	22.8	**31.9**	**0.31**	**46.7**	**33.1**	37.7	0.28	58.6	44.7
CC	47.8	0.21	57.9	30.8	59.1	0.27	57.2	30.4	61.0	0.25	57.1	30.4	61.0	0.20	58.0	31.6

Since a non-homogeneous distribution of scattering objects produces spatially non-homogeneous distribution of – diffuse – reflections, the localization of the noise source at its stationary stage is highly affected by the arrangement of scattering objects. Thus, the sum of the reflections highlights a direction of arrival different to the location where the noise source was physically placed within the scenario. This can be seen when the SNR is 10 dB, and for the case of 90°, where the mode values, strongly affected by the noise source, were different from the true noise source direction, 135°. Therefore, an uneven distribution of scattering objects produced an erroneous localization of stationary sources, strongly affecting to the final results.

5.3. Modified office

This scenario is the same room as the office scenario, but it has been modified by covering the floor and plaster surfaces with carpets and synthetic materials in order to reduce the reverberation time, as described at Section 3.3. This forces to reduce the presence of specular reflections, whereas the diffuse reflections remain the same. Due to the fact that the reverberation times are lower, the relative frequency of the mode was in general higher than in the previous scenario, as can be seen in Table 6. In conditions of ambient noise, and with a high SNR of 30 dB, all the methods except CC reported a good accuracy estimation of the source angle. As expected, performance of CC was lower being the estimation mistaken for the source position of 30°, due to reflections. For lower values of SNR, the results varied in similar way to those of previous scenarios, although the relative frequency of the mode remained higher due to the lower number of reflections.

For a SNR of 10 dB, the mode values did not correspond with the true positions of the voice sources at 60° and 90° for methods PHAT and HT. This is again explained by the presence of the noise source, that can be interpreted as another source of sound more than a noise itself. Besides, at previously pointed out, performance estimation of method CC was not so influenced by the noise increment (see results for SNR of 10 dB). However, its worse accuracy in conditions of ambient noise, high SNR and for the voice source position of 30° (see Table 6) has also lead to conclude that method CC is more vulnerable to reflections and with less accuracy in the angles far from the 90°.

Regarding to the noise, the correlation analysis in this scenario also indicates the existence of a moderate correlation with $\max\{\bar{\varphi}_{n_1 n_2}(\tau)\} = 0.56$, which is similar to that obtained in the

Table 7
Auditorium statistics. The four methods are compared for each combination of the sound source position and SNR. *Mode* is the most probable angle according to the method, *Frequency* is the relative frequency of the mode, *Mean* is the mean value of the observations, and *RMSE* is the root mean square error respect to the real direction of arrival. In this scenario, the noise source is located at 135°. The best method has been highlighted for each situation attending to the mode value and its relative frequency.

Approach	Ambient Noise				SNR = 30 dB				SNR = 20 dB				SNR = 10 dB			
	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE	Mode	Frequency	Mean	RMSE
Angle = 90°																
PHAT	**87.0**	**0.95**	**86.9**	**3.4**	81.0	0.26	91.6	15.6	135.0	0.58	115.7	34.8	136.5	0.98	135.4	45.9
HT	87.0	0.95	86.9	3.4	93.0	0.29	93.2	15.6	135.0	0.47	110.2	32.6	136.5	0.94	133.9	45.3
Roth	87.0	0.91	86.6	3.8	87.0	0.29	90.7	13.6	135.0	0.35	103.9	28.4	136.5	0.86	129.4	43.7
CC	87.0	0.76	88.0	3.5	**87.0**	**0.72**	**87.8**	**4.0**	**87.0**	**0.67**	**88.2**	**4.1**	**90.0**	**0.50**	**88.8**	**6.1**
Angle = 60°																
PHAT	**57.5**	**0.80**	**60.0**	**10.9**	57.5	0.42	68.9	27.9	137.5	0.37	98.0	54.4	137.5	0.84	128.8	73.1
HT	57.5	0.75	59.3	11.7	62.5	0.33	69.8	27.3	137.5	0.37	96.6	53.7	137.5	0.63	119.6	68.0
Roth	57.5	0.74	60.7	12.4	**57.5**	**0.43**	**66.9**	**25.4**	57.5	0.34	76.2	36.9	137.5	0.56	111.3	63.0
CC	62.5	0.27	70.2	15.1	62.5	0.26	70.0	15.2	**62.5**	**0.28**	**70.4**	**15.9**	**62.5**	**0.25**	**71.7**	**18.5**
Angle = 30°																
PHAT	**33.0**	**0.68**	**40.7**	**27.1**	135.0	0.39	73.7	66.9	135.0	0.74	111.5	92.6	136.5	0.94	129.8	103.0
HT	33.0	0.63	40.0	26.6	135.0	0.35	72.7	66.2	135.0	0.69	108.9	91.1	136.5	0.88	125.8	100.6
Roth	33.0	0.65	40.2	25.6	**33.0**	**0.33**	**63.5**	**57.5**	135.0	0.58	94.2	82.2	136.5	0.77	114.4	94.2
CC	45.0	0.19	54.3	32.3	39.0	0.20	54.0	31.6	**39.0**	**0.17**	**54.7**	**33.6**	**37.7**	**0.20**	**55.9**	**35.7**

previous scenario. The noise source suffers same phenomena than in the office scenario, since the scatterers have not been affected by the presence of the new absorbing materials. This example allows to observe how a variation of the reverberation time due to an increase in the absorption coefficient of a scenario, while maintaining the distribution of diffuse reflections, affects the overall detection results, in the sense that the relative frequency of the mode increases.

5.4. Auditorium

The *auditorium* is a scenario with a medium reverberation time (see Section 3.4) even having the biggest volume. This is mainly due to the considerable amount of absorption caused by the materials used in this room. The obtained statistics for this room are presented in Table 7. As can be seen for a SNR of 20 dB, the mode value corresponded with the noise source position for PHAT, HT and Roth methods, independently of the voice source position (except for the Roth method in the case of 60°). Even for a SNR of 30 dB, the same phenomena occured for PHAT and HT when the voice source was located at 30°. The difference between this case and the previous scenarios should be noted. Previously, this phenomena was significant only for lower SNRs, whereas it is now considerable in most of the source position/SNR ratio combinations evaluated. It is worthy to note, that for these cases, the relative frequency of the mode increased as the SNR decreased, and the mean value is nearer the noise position, indicating that methods are treating the noise as another voice source. In contrast, the CC method is the only one that estimates the voice source according to its mode value, even with a SNR of 10 dB. It can be noted that,

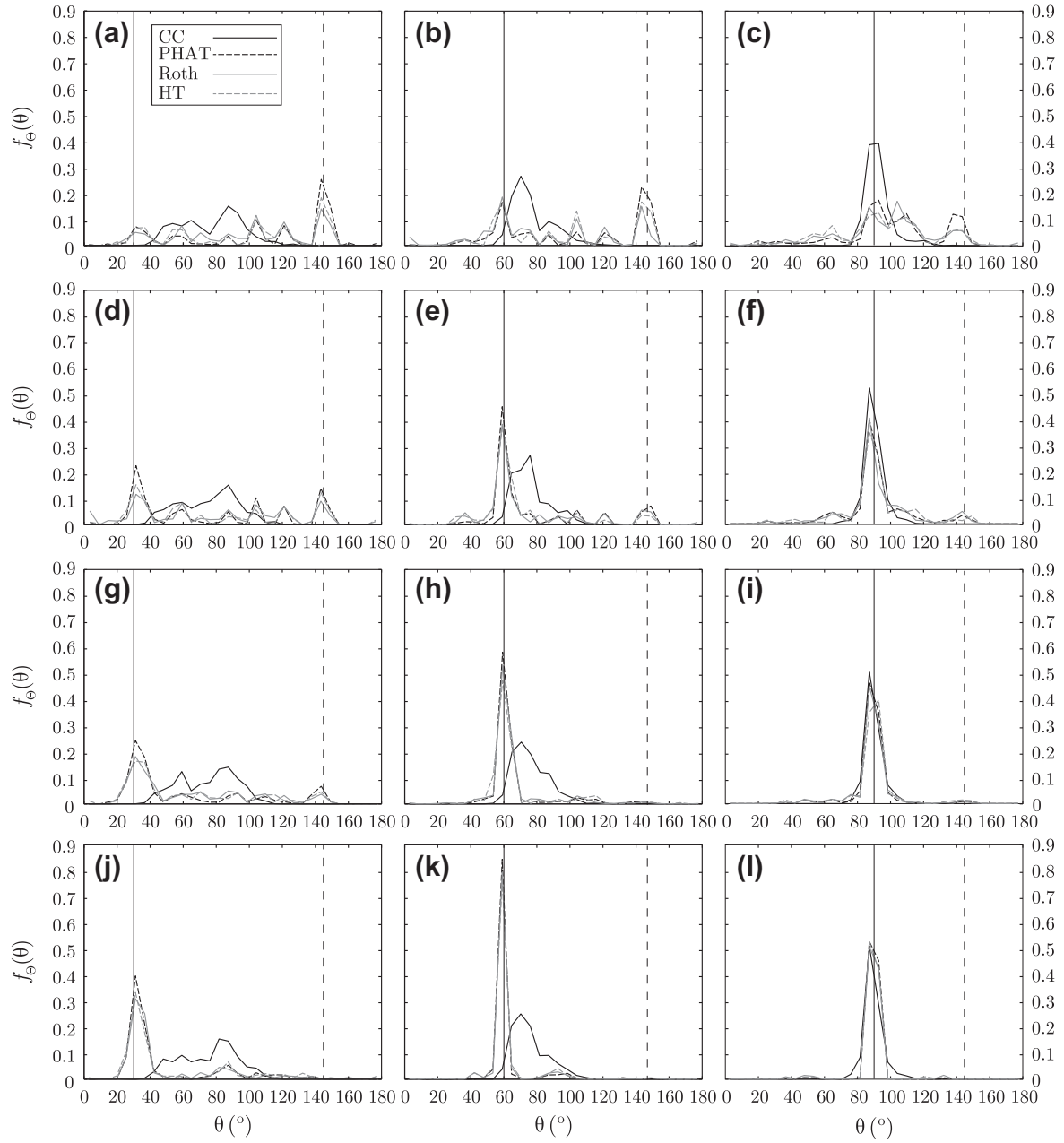

Fig. 6. Results obtained in lecture room scenario. Each subfigure contains the superimposed histograms of the four methods given a location of the source of sound and a signal-to-noise ratio. The signal-to-noise ratio is 10 dB from (a) to (c), 20 dB from (d) to (f), 30 dB from (g) to (i) and with ambient noise from (j) to (l). The source of sound (vertical line) is located at 30°, 60°, and 90° in the first, second and third column, respectively. The noise source (vertical dashed line) is located at 145°.

despite of being CC less robust to reverberation effects, it estimated with more accuracy voice source positions mistaken by the other methods, since the latter take the noise source as the localization target.

Previously described effects on performance results of the acoustical properties of this room are mainly due to the presence of very few significant reflections – high absorbing walls, causing highly correlated noise signals arriving to both microphones. In order to assess the extent to which this noise was correlated, a normalized cross-correlation was also obtained as in previous scenarios, reporting the highest value, with $\max\{\bar{\varphi}_{n_1 n_2}(\tau)\} = 0.78$. This level of noise spatial correlation affected considerably PHAT, Roth and HT methods since they assume noise has to be uncorrelated. Therefore, these methods detected the noise source as a secondary voice source, causing an important source of inaccuracies. Under these conditions, the noise source was detected clearly as another source with a higher influence in the estimation results, since the noise was always emitting whereas voice was not.

This is explained by the high absorption of the room, since the reflected waves arrive highly attenuated and only the direct path of the noise contributes. It has to be highlighted how, in this scenario, the reverberation time is similar to the office; however, the absorbing properties of the room are clearly different, providing different results. This leads to conclude that not only reverberation time and SNR should be considered in the validation of TDOA algorithms, but also the overall absorbing/diffracting properties of boundaries at the room. An alternative way to address this conclusion is through the value of the relative frequency of the mode when the mode points towards the noise source position. In conditions

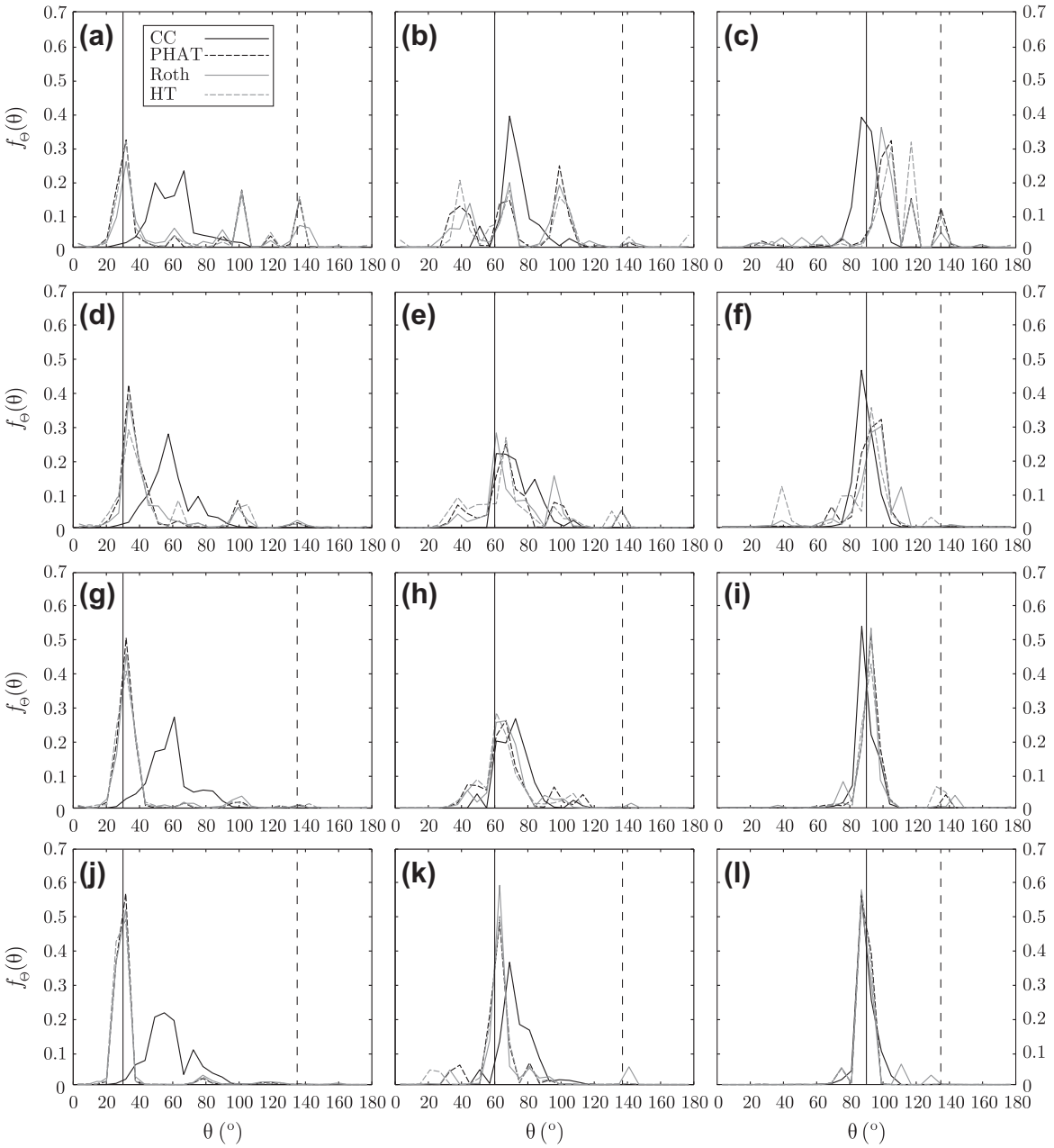

Fig. 7. Results obtained in office scenario. Each subfigure contains the superimposed histograms of the four methods given a location of the source of sound and a signal-to-noise ratio. The signal-to-noise ratio is 10 dB from (a) to (c), 20 dB from (d) to (f), 30 dB from (g) to (i) and with ambient noise from (j) to (l). The source of sound (vertical line) is located at 30°, 60°, and 90° in the first, second and third column, respectively. The noise source (vertical dashed line) is located at 135°.

with a SNR of 10 dB, frequency values of the mode are much higher than those obtained in the other scenarios. Indeed, they were very close to the ideal value of 1.0, which clearly indicates the strong detection at the precise direction of the noise source.

Comparing this scenario with the rest when the SNR is 10 dB, it can be seen that these values are much higher, sometimes close to the ideal value of 1.0, which clearly indicates the strong detection at the precise direction of the noise source.

As a conclusion, the average absorption coefficient and diffusing properties of walls have an important role in the accuracy of these methods, even more than just taking into account the reverberation time. Therefore, when one sound source and one highly spatial correlated noise are present in the room, clearly method CC is the more advantageous algorithm in conditions of moderate SNR, since the other algorithms will estimate the noise source position as the voice one.

5.5. General discussion

The experimental results described in the previous sections about the behavior of TDOA methods in different real scenarios have allowed characterizing these methods in terms of performance and addressing the influence of acoustic features not typically analysed.

Previous studies have stated that the performance of these methods is usually worse under conditions of low SNR. However, the most important effect outlined by this study is that the influence of noise in the arrival direction estimation depends strongly

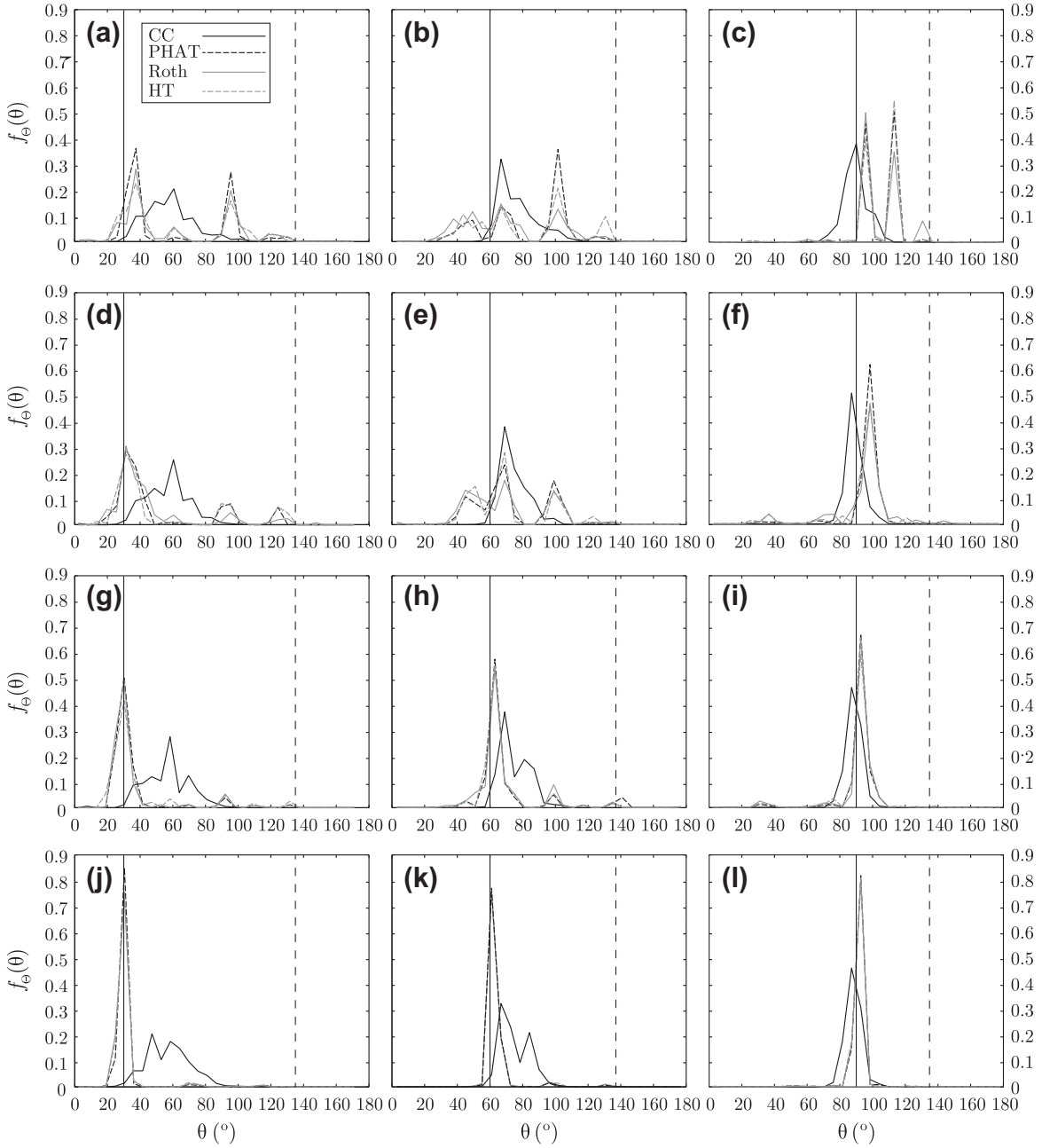

Fig. 8. Results obtained in modified office scenario. Each subfigure contains the superimposed histograms of the four methods given a location of the source of sound and a signal-to-noise ratio. The signal-to-noise ratio is 10 dB from (a) to (c), 20 dB from (d) to (f), 30 dB from (g) to (i) and with ambient noise from (j) to (l). The source of sound (vertical line) is located at 30°, 60° and 90° in the first, second and third column, respectively. The noise source (vertical dashed line) is located at 135°.

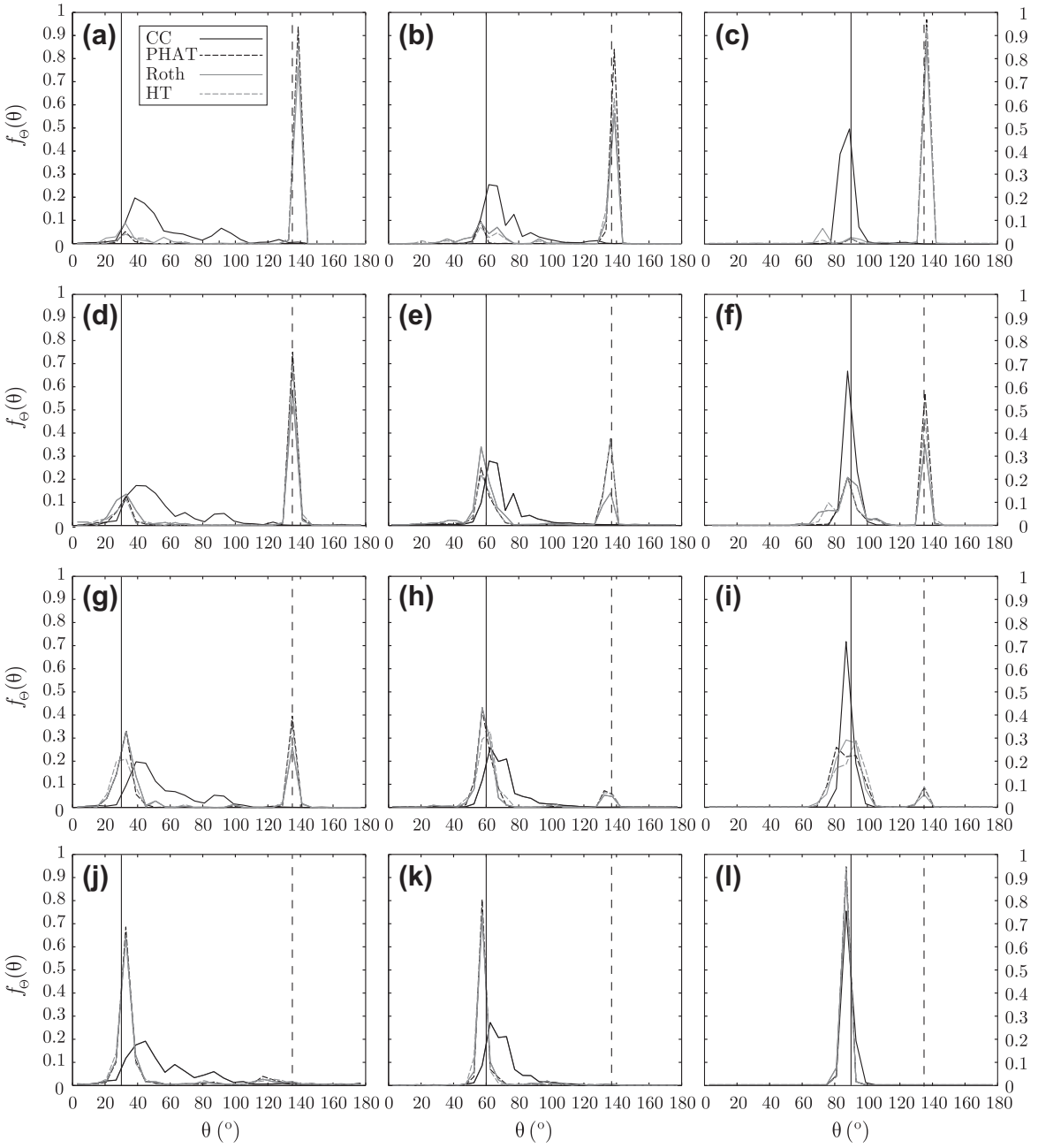

Fig. 9. Results obtained in auditorium scenario. Each subfigure contains the superimposed histograms of the four methods given a location of the source of sound and a signal-to-noise ratio. The signal-to-noise ratio is 10 dB from (a) to (c), 20 dB from (d) to (f), 30 dB from (g) to (i) and with ambient noise from (j) to (l). The source of sound (vertical line) is located at 30°, 60°, and 90° in the first, second and third column, respectively. The noise source (vertical dashed line) is located at 135°.

on the features of the room under test. This analysis has also led to conclude, that noise cannot be assumed uncorrelated in non-simulated environments. In these scenarios, the correlation between noise signals depends exclusively on the room characteristics. Thus, the noise recorded at two microphones becomes more uncorrelated with an increase of the existence of scattering/reflecting objects. In a general way, the results obtained in the different experimental conditions have shown that the room features have affected the level of noise correlation, which in turn, has affected the overall performance estimation of these methods, as they are based on the assumption of uncorrelation. This effect has been characterized in terms of performance degradation. In particular, the noise has influenced in a displacement of the mean value far from the mode and an increase of the RMSE. Furthermore, the mode has been shifted by the noise in low SNRs situations, which affects negatively the estimation.

Furthermore, a higher probability has also been found for the arrival directions associated to the noise source in scenarios with higher absorption. Thus, as expected, a higher absorption increases the level of correlation between the signals recorded at the two microphones. In these conditions, TDOA methods analysis must deal with a noise source as if it was an additional voice source, being more negative its effect in conditions of low SNR. The results obtained have also shown that the simple CC method with no

weighting function, seems to be more suitable for scenarios where noise cannot be assumed uncorrelated, although being more vulnerable to reverberance.

Usually, in the literature, TDOA methods that modify method CC by introducing a weighting function have been proved to be more accurate with high SNR and under low reverberation conditions [3]. The study presented here has highlighted that noise features and room acoustic characteristics should be considered to establish this kind of assessment. Thus, the modification of the noise due to the acoustic properties of the scenario should be considered before analysing its effect, because it can act as a source of interest for the estimation method if the correlation is high enough.

Multiple reflections in the environment lead to an increase of the RMSE. Traditionally, this performance degradation of methods has been associated to reverberation time. The comparison carried out between two scenarios with very different reverberation times – office and lecture room – has also shown this. The negative influence in the performance of TDOA methods has been very evident in these scenarios. In particular, the comparison performed among methods has highlighted method CC as the most vulnerable to multiple reflections and PHAT as the least affected. As a step forward, the comparison carried out between scenarios with similar reverberation times but with different overall absorbing/diffusing properties – i.e. office and auditorium – has also highlighted the necessity of considering other acoustic properties of the scenarios. In particular, the influence of absorbing/diffusing properties, modified by the number of scattering objects existing in the room, affects performance by changing the correlation property between the noise signals acquired in the microphones. Thus, in a scenario with high diffusing materials, the sound captured by microphones can be considered partially uncorrelated and therefore, methods PHAT, Roth and HT estimate properly the voice direction if the SNR is not low. In a scenario with a high spatial correlation of the noise captured in the microphones, such as the auditorium, CC is the most accurate method, since voice and noise are very correlated and the method detects the source with more energy level.

In general, weighting functions introduced in cross-correlation algorithms make the methods more robust against the effect of multiple reflections. However, if the number and energy of reflections increase, these methods become more inaccurate. Under low and medium reverberation times PHAT, Roth and HT methods are similar in their performance, being Roth slightly more robust against correlated noise, while PHAT stands out under high reverberation times. Also, related to the detection accuracy, when the sound source is placed in front of the microphones, all the methods are more robust in the estimation, whereas the performance gradually decreases as these angles approximate the axis that contains both microphones (azimuth values of 0° and 180°).

To summarize, the average absorption coefficient and diffusing properties of walls have an important role in the performance of these methods, even higher than the influence of reverberation time in presence of an external noise source.

6. Conclusion

Within the field of direction of arrival estimation, most of the previous studies have just dealt with simulated and ideal scenarios, such as empty rooms. Thus, conclusions have been outlined without considering the overall effects associated with the reflection phenomena – specular or diffuse reflections – and, in many cases, upon the assumption of idealized – uncorrelated – noise sources. Due to this lack of generalized cross-correlation methods or TDOA methods validation in adverse scenarios, one of the main issues raised in this study has been the performance evaluation of four TDOA methods for the localization of a voice sound in real scenarios acoustically characterized and under the presence of a controlled noise source.

The experiments validate most of the conclusions already formulated based on room acoustic simulations, regarding the dependence of these methods performance on the reverberation time and SNR. However, new conclusions can be formulated based on the results obtained from real scenarios, highlighting the dependence of the methods with the overall nature of reflections. The spatial correlation of the noise at the measuring positions strongly depends on the amount of specular or diffuse reflections occurring. Thus, a scenario characterized for having a low number of diffuse reflections may alter the nature of the noise captured at different positions, by making it more correlated. The influence of the noise nature in TDOA performance must then be considered, by addressing its degree of correlation, which is not related to reverberation times but to the amount of diffusive reflections.

Regarding to the particular performance of algorithms, PHAT, Roth and HT usually have been highlighted as the most accurate methods for general purpose TDOA applications, even with high reverberance and low signal-to-noise ratios. However, empirical results presented here have also indicated that their performance is strongly affected by the noise when the scenario is built with high absorbing and poor diffusive materials. The empirical evaluation has also outlined the better performance of method CC with spatially correlated noise detected in two microphones, because the other methods can estimate the noise as another voice source. On the other hand, method CC is the weakest algorithm under conditions of multiple and strong reflections, and methods PHAT, Roth and HT show a similar performance for medium levels of reverberation times. The main difference among these methods performance arises for high values of reverberation, being PHAT the most robust method when the signal-to-noise ratio is moderate, validating previously published results, and Roth the most tolerant to correlated noise when the SNR is low.

To conclude, the validation of TDOA algorithms are usually based on signal-to-noise ratio and reverberation times. However, the experiments of this study suggest that, for a better validation of TDOA algorithms in realistic environments, an architectural acoustic point of view should be considered, specifically considering the extent to which they alter the correlation properties of the noise registered in the microphones. Clearly, the use of additional acoustic parameters that have not been reported in the literature is an interesting line of research. Future work should be focused on a more detailed study on how the overall boundary conditions, i.e. absorption and diffusion, have a decisive influence in the performance of GCC methods, as well as the inclusion of other information about the acoustical behaviors of the scenarios in the evaluation of arrival estimation methods.

Acknowledgements

This work has been supported by the University of Jaén and Caja Rural de Jaén (Spain) under the Project 2009/12/12 and the Science and Innovation Department of the Spanish Government under the project TIN2011-27512-C05-01.

Appendix A. Histograms

In this section the histograms representing the probability function $f_{\Theta}(\theta)$ obtained for each scenario are depicted. Each subfigure represents a single situation, where the sound source is placed at a specific angle for a given SNR. The histograms of the four methods are superimposed. A vertical line represents the localization of

the voice source, while a vertical dashed line represents the localization of the noise source (see Figs. 6–9).

References

[1] Havelock D, Kuwano S, Vorländer M. Handbook of signal processing in acoustics. Springer; 2008.
[2] Marković I, Petrović I. Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering. J Robot Auton Syst 2010;58(11):1185–96.
[3] Martin R, Heute U, Antweiler C. Acoustic source localization with microphone arrays, in advances in digital speech transmission. John Wiley & Sons; 2008.
[4] Cobos M, Lopez JJ, Martinez D. Two-microphone multi-speaker localization based on a Laplacian mixture model. Digital Signal Process 2011;21(1):66–76.
[5] Cobos M, Marti A, Lopez JJ. A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Process Lett 2011;18(1):71–4.
[6] Ferreira JF, Pinho C, Dias J. Implementation and calibration of a Bayesian binaural system for 3D localisation. In: IEEE international conference on robotics and biomimetics; 2008. p. 1722–7.
[7] Morrissey RP, Ward J, DiMarzio N, Jarvis S, Moretti DJ. Passive acoustic detection and localization of sperm whales (Physeter macrocephalus) in the tongue of the ocean. Appl Acoust 2006;67(11–12):1091–105.
[8] Giraudet P, Glotin H. Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely-spaced hydrophone array. Appl Acoust 2006;67(11–12):1106–17.
[9] Tiana-Roig E, Jacobsen F, Fernandez Grande E. Beamforming with a circular microphone array for localization of environmental noise sources. J Acoust Soc Am 2010;128(6):3535–42.
[10] Murray JC, Erwin H, Wermter S. Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In: Proceedings of neurobotics workshop. Ulm, Germany; 2004. p. 89–97.
[11] Trifa VM, Koene A, Moren J, Cheng G. Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In: The 16th IEEE international symposium on robot and human interactive communication; 2007. p. 393–8.
[12] May T, van de Par S, Kohlrausch A. A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans Audio Speech Lang Process 2011;19(1):1–13.
[13] Faller C, Merimaa J. Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J Acoust Soc Am 2004;116(5):3075–89.
[14] Zhang W, Rao B. A two microphone-based approach for source localization of multiple speech sources. IEEE Trans Audio Speech Lang Process 2011;18(8):1913–28.
[15] Cobos M, Lopez JJ. Two-microphone separation of speech mixtures based on interclass variance maximization. J Acoust Soc Am 2010;127(3):1661–72.
[16] Knapp C, Carter G. The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process 1976;24(4):320–7.
[17] Bartsch C, Volgenandt A, Rohdenburg T, Bitzer J. Evaluation of different microphone arrays and localization algorithms in the context of ambient assisted living. In: International workshop on acoustic echo and noise control; 2010.
[18] Brutti A, Omologo M, Svaizer P. Comparison between different sound source localization techniques based on a real data collection. Hands-Free Speech Commun Microph Arrays 2008.
[19] Marti A, Cobos M, Lopez JJ. Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2011. p. 2592–5.
[20] Mungamuru B, Aarabi P. Enhanced sound localization. IEEE Trans Syst Man Cybern Part B: Cybern 2004;34(3):1526–40.
[21] Omologo M, Svaizer P. Use of the crosspower-spectrum phase in acoustic event location. IEEE Trans Speech Audio Process 1997;5(3):288–92.
[22] Ui-Hyun K, Jinsung K, Doik K, Hyogon K, Bum-Jae Y. Speaker localization on a humanoid robot's head using the TDOA-based Feature Matrix. In: IEEE international symposium on robot and human interactive communication (RO-MAN); 2008.
[23] Valin JM, Michaud F, Rouat J, Lêtourneau L. Robust sound source localization using a microphone array on a mobile robot. In: International conference on intelligent robots and systems; 2003. p. 1228–33.
[24] Wang H, Chu P. Voice source localization for automatic camera pointing system in videoconferencing. In: IEEE international conference on acoustics, speech, and signal processing; 1997. p. 187–90.
[25] Yu Y, Silverman HF. An improved TDOA-based location estimation algorithm for large-aperture microphone arrays. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2004. p. 77–80.
[26] Zhang C, Florêncio D, Zhang Z. Why does phat work well in low noise, reverberative environments? In: IEEE international conference on acoustics, speech and signal processing; 2008. p. 2565–8.
[27] Benesty J. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J Acoust Soc Am 2000;107(1):384–91.
[28] Rui D, Florencio D. Time delay estimation in the presence of correlated noise and reverberation. IEEE Int Conf Acoust Speech Signal Process 2004.
[29] Yang J, Lee C, Kim S, Kang H. A robust time difference of arrival estimator in reverberant environments. In: 17th European signal processing conference; 2009. p. 864–8.
[30] Chen J, Benesty J, Huang Y. Time delay estimation in room acoustic environments: an overview. Eurasip J Appl Signal Process 2006.
[31] Koopmans L. The spectral analysis of time series. New York: Academic; 1974.
[32] Carter G, Nuttall A, Cable P. The smoothed coherence transform (SCOT). Naval Underwater Systems Center, New London Lab.; 1972. Tech. Memo TC-159-72.
[33] Roth P. Effective measurements using digital signal analysis. IEEE Spect 1971;8(4):62–70.
[34] Hannan E, Thomson P. Estimating group delay. Biometrika 1973;60(2):241–53.
[35] Gustaffson T, Rao B, Trivedi M. Source localization in reverberant environments: modeling and statistical analysis. IEEE Trans Speech Audio Process 2003;11:791–803.
[36] Hodgson M. When is diffuse-field theory applicable? Appl Acoust 1966;49(3):197–207.
[37] Cook RK, Waterhouse RV, Berendt RD, Edelman S, Thompson MC. Measurement of correlation coefficients in reverberant sound fields. J Acoust Soc Am 1955;27(6):1072–7.
[38] Jacobsen F. The diffuse sound field. The Acoustic Laboratory, Technical University of Denmark, 1979. Report no. 27.
[39] ISO/DIS 3382. Acoustics – measurement of the reverberation time of rooms with reference to other acoustical parameters; 1997.
[40] Kuttruff H. Room acoustics. 4th ed. Taylor & Francis; 2000.
[41] Barron M. Interpretation of early decay times in concert auditoria. Acustica 1995;81:320–31.
[42] Vincent E, Sawada H, Bofill P, Makino S, Rosca JP. First stereo audio source separation evaluation campaign: data, algorithms, and results. International conference on independent component analysis and signal separation (ICA 2007) 2007.
[43] Montgomery DC, Runger GC. Applied statistics and probability for engineers. John Wiley & Sons; 2011.
[44] Rife D, Vanderkooy J. Transfer-function measurements using maximum-length sequences. J Audio Eng Soc 1989;37:419–44.
[45] Farina A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Audio Engineering Society 108th Convention. Paris, France; 2001.
[46] Schroeder MR. New method of measuring reverberation time. J Acoust Soc Am 1965;37:409–12.
[47] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 1979;65:943–50.