Observation of Vehicle Axles through Pass-by Noise: a Strategy of Microphone Array Design

Patrick Marmaroli*, Mikael Carmona†, Jean-Marc Odobez‡, Xavier Falourd*, Hervé Lissek*
*Laboratory of Electromagnetism and Acoustics, EPFL, Lausanne, Switzerland, email: patrick.marmaroli@epfl.ch
†CEA / DRT/DSIS//LCFE, CEA-Léti DSIS, Grenoble, France
‡Idiap Research Institute, Martigny, Switzerland

***Abstract*—This paper focuses on road traffic monitoring using sounds and proposes, more specifically, a microphone array design methodology for observing vehicle trajectory from acoustic-based correlation functions. In a former work, authors have shown that combining generalized cross-correlation (GCC) functions and a particle filter (PF) onto the audio signals acquired simultaneously by two sensors placed near the road allows the joint estimation of speed and wheelbase length of road vehicles as they pass-by. This is mainly due to the broadband nature of the tyre/road noise which makes their spatial dissociation possible by means of an appropriate GCC processor. At the time, nothing has been said about the best distance to chose between the sensors. A methodology is proposed here to find this optimum, which is expected to improve the observation quality and thus the tracking performance. Theoretical developments of this paper are partially assessed with preliminary experiments.**

I. Introduction

ROAD traffic monitoring (RTM) plays a key role in ensuring road safety, predicting traffic jam, measuring noise, assessing environmental impact on urban areas etc. Amongst existing techniques, passive acoustic ones present the advantage of being non-intrusive, health safety (no wave emission) and multi-use, *i.e.* different kinds of information may be extracted from the same observation, depending on the associated signal processing algorithm. That is why a large community of acoustic/signal processing researchers are working on the challenge of equalling, or even outperforming, the performance of active or intrusive technologies, based on the power of modern-day computing.

Since the mid 1990s, more and more attention has been paid to passive acoustic-based systems for traffic monitoring. In 1996, vehicle classification using wavelet decomposition of audio signals were investigated by Choe *et. al* in [1]. In 1997, Chen *et al.* [2] and Forren *et al.* [3] independently investigated the detection problem using cross-correlation functions between sensors spatially disjoint. The counting problem was also handled by Brockman *et al.* in 1997 [4] and Kuhn *et. al* [5] in 1998 which respectively deployed an auto-regressive algorithm based on a pass-by spectrum model (one sensor) and a beamforming-based technique (80 sensors) to detect vehicle presence. Other kind of counting techniques based on correlation and filters have emerged later [6]–[10]. The speed estimation problem has also been addressed extensively, for instance in [11]–[16]. A recent trend consists in considering the pass-by noise as a measure of the energy consumption: in 2011, Can *et al.* successfully showed the correlation between emitted airborne pollutant and road traffic noise near a highway [17]. This is a brief overview of what information traffic noise can provide.

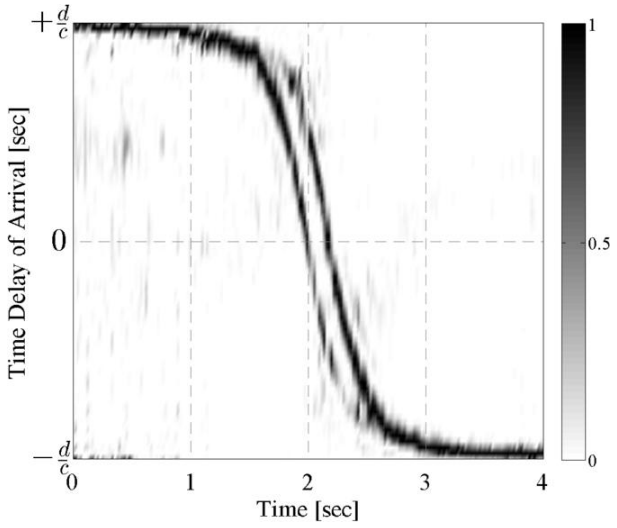


Fig. 1. Typical CCTS of a road vehicle pass-by (about 50 km/h). d is the inter-sensor distance and c is the speed of sound.

In this paper, we are interested in observing and estimating the wheelbase length of road vehicles using pass-by noise. The wheelbase estimation problem has been rarely addressed in the acoustic literature. Yet, it is an important feature for vehicle classification. In [16], Cevher *et. al* suggested a wave patterns-based recognition algorithm enabling the joint speed and wheelbase estimation from a one-channel pass-by recording acquired on the roadside. Engine, tyre, exhaust and air turbulence noises are meticulously modeled but presence of interfering noises in the monitored area may limit its applicability. In a former work, [18], we opted for a two microphone array-based procedure. The idea was to concatenate successive cross-correlation measurements and apply a particle filter (PF) to the obtained image where the position of each axle and their common speed were included in the state of the target. As an example of observation, Fig.1 depicts what we called a Cross-Correlation Time Series (CCTS) of a vehicle pass-by (nearly 50 km/h) with two dimensions: TDOA versus Time. TDOA is bounded by $\pm d/c$ where d is the inter-sensor distance and c is the speed of sound. Two traces, one per axle, are clearly distinct when

the vehicle is in front of the array (*i.e.* TDOA = 0 ms). The slope of both traces is directly related to the vehicle speed, and their space is directly related to the wheelbase length. In [18], authors shown the promising results of applying a Bayesian filter on such an observation, especially for cases where multiple vehicles pass each other in front of the two-element array. But at the time, nothing was said about the microphone array aperture which need to be meticulously adjusted to provide the best observation (CCTS) as possible. Mathematically speaking, it is a well-known result that the optimal microphone arrangement for TDOA-based sound source localization is the Platonic-shaped array surrounding the target [19], [20]. In the RTM context, such a geometry is difficult, if not impossible, to achieve. In the present case, we are looking for the optimal d (distance between the two sensors of the array) for which the two traces inherent to the rear and front axles in the CCTS are clearly depicted.

The remainder of this paper is structured as follows. In section II, the objective and theoretical background of sound source localization since time delay estimates are introduced. In section III, methods are proposed for finding minimal, maximal and optimal inter-sensor distances. Preliminary experimental results are discussed in section IV. A final discussion concludes the article in section V.

II. Problem definition

Let us consider the scenario depicted in Fig.2. A two-element microphone array with inter-sensor distance d is placed on the roadside. Both microphones are placed at the same distance D from the road lane. Road vehicles are modeled as two stochastic and identically distributed processes separated by the wheelbase length w_b. The distance between the closest point of approach (CPA) of the vehicle and the front axle is denoted x_0. Both axles are also identified by their respective direction of arrival (DOA) θ_1 and θ_2 on the array.

A commonly accepted approximation consists in saying that the mechanical noise predominates for vehicles running at low speed (below 50 km/h) and the tyre/road noise predominates for vehicles running at higher speeds. But over time, more and more modern cars make it the tyre/road noise always dominates even in congested urban situation for constant speed driving [21]. The model of Fig.2 seems therefore reasonable for a wide scope of scenario. In this paper, speed is simply assumed to be a constant during the observation (nearly one second). We should also mention that one observation results from the concatenation of successive 30 ms audio frames processing during which the vehicle is considered as static.

A. Signal model

Let N be the number of zero mean, broadband and uncorrelated sound sources located at coordinate $\mathbf{r}_n^s$, $1 \leq n \leq N$. Let x_1 and x_2 be the audio signal acquired by both microphones located at coordinate $\mathbf{r}_1^m$ and $\mathbf{r}_2^m$ respectively. Without loss

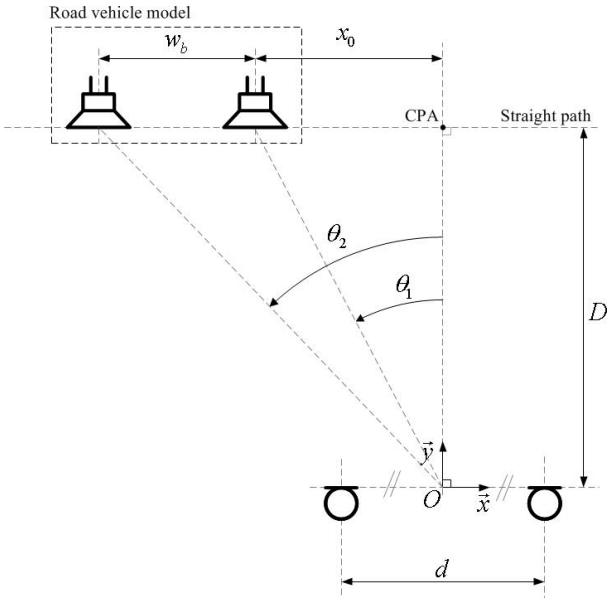


Fig. 2. Bimodal sound source model of a two-axle road vehicle , wavefronts are acquired by a microphone array placed in parallel to the road lane at a distance D from the vehicle closest point of approach (CPA).

of generality, sensor 1 is taken as the reference microphone. Assuming an ideal free field, homogeneous medium of propagation and no energy loss between the two sensors, x_1 and x_2 can be modeled as:

$$x_1(t) = \sum_{n=1}^{N} s_n(t-\delta_n) + w_1(t), \quad (1)$$

$$x_2(t) = \sum_{n=1}^{N} s_n(t-\delta_n-\tau_n) + w_2(t), \quad (2)$$

where δ_n is the time of flight between the n^{th} source and the reference microphone, w_m is an additive measurement noise, considered as a wideband, stationary, zero-mean Gaussian process, uncorrelated both with the signals and noise at other sensors, and τ_n is the TDOA between both sensors of the n^{th} incoming wavefront.

In the considered applied framework, the model (1)-(2) is restricted to N=2 where $s_1(t)$ and $s_2(t)$ are supposed to be the sound produced by front and rear tyre-asphalt interactions respectively. Under far-field assumption, the wheelbase length w_b is related to sound sources TDOA's by:

$$w_b = D\left(\tan\theta_2 - \tan\theta_1\right), \quad (3)$$

with

$$\theta_i = \arcsin\left(\frac{c\tau_i}{d}\right) \ \mathrm{i} \in [1,2], \quad (4)$$

and where c is the speed of sound. After (3) and assuming that D and d are known, the wheelbase length estimation problem is turned into a time-delay estimation problem. The time-delay estimator on which we rely on is presented in the next section.

B. Time-delay estimation

It is a well-known result that in presence of a single broadband source, *i.e.* $N = 1$ in (1)-(2), the optimal estimator of τ_1 is the lag corresponding to the maximum value of the cross-correlation between $x_1(t)$ and $x_2(t)$ [22]. In that case, one can also give an explicit expression of the Cramer-Rao lower bound (CRLB), which depends on the spectral bandwidth of the source and on the signal-to-noise ratio. If $N > 1$, the optimal estimator cannot be computed if the sources spectrum are not exactly known. Consequently, two strategies can be considered: undertake a source identification process (requiring a high number of sensors to achieve a spatial filtering for instance) or derive a suboptimal estimator which will process directly the observations, considering the signal-to-noise ratio is high enough. As the proposed approach implies two microphones only, we relied on the traditional generalized cross-correlation (GCC) functions which are suboptimal time delay estimators but very popular for their robustness and weak computation requirements. They are expressed by [23]:

$$R^g_{s_1s_2}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f) X_1(f) X_2^*(f) e^{2j\pi f\tau} df, \qquad (5)$$

where $(.)^*$ stands for the complex conjugate operator, f denotes the frequency (Hz), $X_1(f)$ (respectively $X_2(f)$) is the Fourier transform of $x_1(t)$ (respectively $x_2(t)$), τ stands for the time lag and $\psi_g(f)$ is a weighting function. For instance, setting $\psi_g(f) = 1\ \forall f$ turns the expression (5) into the classical cross-correlation function. In the single source case, an estimation of the TDOA τ_1 is given by looking at the argument of the peak value of the GCC function:

$$\hat{\tau}_1 = \arg\max_{\tau} R^g_{s_1s_2}(\tau) \qquad (6)$$

In order to accentuate the peak, different weighting functions were investigated in the literature regarding the acoustical conditions. In the sound source localization community, one of the most successful processor is The PHAse Transform weighting (PHAT). It is expressed by [23]:

$$\psi_{phat}(f) = \begin{cases} \frac{1}{|X_1(f)X_2^*(f)|} & \text{if } |X_1(f)X_2^*(f)| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \qquad (7)$$

Heuristically developed in the middle of the 1970s, the GCC-PHAT function proved to perform very well under realistic acoustical conditions. Reasons for its success are numerous: its implementation is straightforward, no *a priori* knowledge on signal and noise is required, it is more consistent than some other GCC members when the characteristics of the source change over time [24]. Also it was recently proved that in case of high signal to noise ratio, the GCC-PHAT function is the optimal time-delay estimator in a maximum likelihood sense, regardless of the amount of reverberation [25]. Besides, many comparative studies proved its robustness in presence of multipath distorsion, see for instance [26], [27].

After (5)-(7), the PHAT processor may be seen as a crosspower spectrum whitening [28] discarding any magnitude information contained in the audio signals. That makes it well adapted to cases where pairwise amplitude differences cannot be used as a relevant feature for localization, typically, when the microphone array has a small aperture in comparison with the distance to the source. But the main problem is that any spatially coherent noise, even when lower than the signal of interest, results in a spurious peak in the PHAT correlation function. Unfortunately, such a kind of noise may be frequent in outdoor monitoring (industrial/agricultural noises, birds, pedestrian activity etc.). One way to overcome this problem is to apply the PHAT transform on a pre-defined spectral band only. This is achieved using the *Bandpass-PHAT* (BPHAT) weighting. This processor was previously proposed for speaker localization by DiBiase in [29] p. 46 or for water pipes leaks localization by Gao *et al* in [30], [31]. It is defined as:

$$\psi_{bphat}(f) = \begin{cases} \psi_{phat}(f) & \text{if } f_c - B_w/2 \leq |f| \leq f_c + B_w/2 \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

where f_c and B_w respectively denote the central frequency and the bandwidth on which the BPHAT transform is applied. To be effective, the spectral band on which the BPHAT is applied needs to be identical or within the bandwidth of the signal of interest. For the sake of simplicity, the perfect equality is assumed in this paper. According to (8) and (5), one can demonstrates that the closed-form expression of the GCC-BPHAT function for the single source case is (see Appendix A):

$$R^{bphat}_{s_1s_1}(\tau) = 2B_w \cos\left[2\pi f_c(\tau - \tau_1)\right] \operatorname{sinc}\left[B_w(\tau - \tau_1)\right]. \qquad (9)$$

For the two sound sources case and under the assumption that each source delivers a zero-mean signal, uncorrelated with the other, one gets:

$$\begin{aligned} R^{bphat}_{s_1s_2}(\tau) &= R^{bphat}_{s_1s_1}(\tau) + R^{bphat}_{s_2s_2}(\tau), \qquad (10) \\ &= 2B_w\left(A_1 + A_2\right), \qquad (11) \end{aligned}$$

with

$$A_k = \cos\left[2\pi f_c(\tau - \tau_k)\right] \operatorname{sinc}\left[B_w(\tau - \tau_k)\right],\ k \in [1, 2].$$

It may be noted that, regarding the application targeted for these developments, the non-correlation of the two sources is a debatable assumption. Sounds come from the axles of the same vehicle and would somewhat be correlated (mechanical/structured connection between axles, identical speed and loading etc.). But as a first approximation, cross-terms in the correlation measure are neglected in this paper.

III. Inter-sensor distance optimization

After Eq. (11), the characteristics of the peaks (width, emergence and spacing) depend on the spectral properties of the sources (B_w, f_c) and the geometrical parameters of the scene (x_0, D, w_b, d). *In-situ*, distances D and d are the only adjustable parameters, except for normative measurements where D is imposed. The challenge therefore consists in finding the optimal d ensuring the best observation of the two traces in the BPHAT-CCTS. This is what is addressed in the following.

A. Cramer-Rao Lower Bound

The Cramer-Rao Lower Bound (CRLB) defines the smallest variance than can be achieved by an unbiased estimator. It is based on the Fisher information matrix. For cases when the estimator depends on variable parameters, the CRLB enables their optimization.

Suppose that the parameter to estimate is w_b and the parameter to optimize is d. The available measurements are $\tau_{12,1}$ and $\tau_{12,2}$, simply noted τ_1 and τ_2 below, such that:

$$\tau_1 = \hat{\tau}_1 + n_1, \quad (12)$$

$$\tau_2 = \hat{\tau}_2 + n_2, \quad (13)$$

where $\hat{\tau}_j$ is an estimate of τ_j and n_j is a zero mean gaussian noise with variance σ_j^2 denoting the uncertainty on the measurement, $j \in [1, 2]$. $\hat{\tau}_2$ can be expressed as a function of $\hat{\tau}_1$ and w_b:

$$\hat{\tau}_2 = f(\hat{\tau}_1, w_b). \quad (14)$$

After Eq.(3), it comes:

$$\theta_2 = arctan\left(\frac{w_b}{D} + tan\theta_1\right) \quad (15)$$

Replacing θ_1 and θ_2 by their expressions in (4) yields f:

$$f(\hat{\tau}_1, w_b) = \frac{d}{c}\sin\left\{\arctan\left[\tan\left(\arcsin\left(\frac{c\hat{\tau}_1}{d}\right)\right) + \frac{w_b}{D}\right]\right\}. \quad (16)$$

The CRLB is defined as the inverse of the Fisher information matrix. The latter is given by [32] page 47:

$$F = A'\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} A, \quad (17)$$

where

$$A = \begin{pmatrix} \partial\tau_1/\partial\hat{\tau}_1 & \partial\tau_1/\partial w_b \\ \partial\tau_2/\partial\hat{\tau}_1 & \partial\tau_2/\partial w_b \end{pmatrix}, \quad (18)$$

$$= \begin{pmatrix} 1 & 0 \\ \partial f/\partial\hat{\tau}_1 & \partial f/\partial w_b \end{pmatrix}. \quad (19)$$

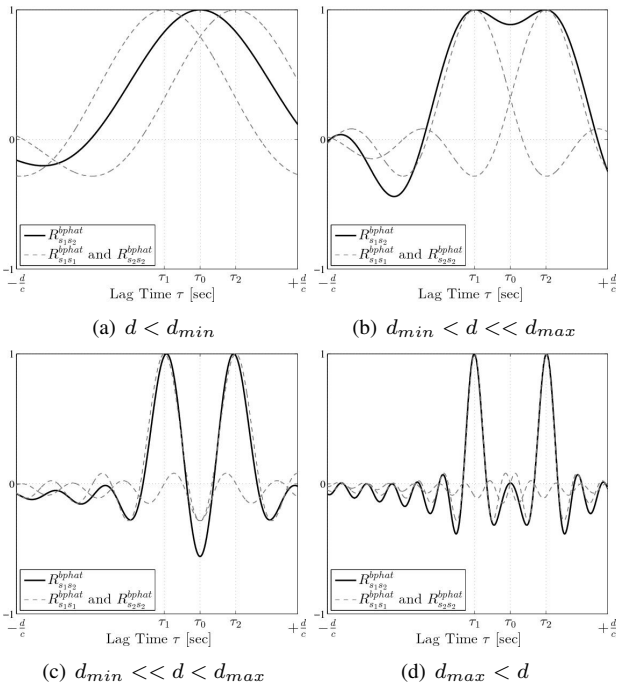
The optimal d is the one which maximizes the determinant of F (D-optimality criterion) [33]. The determinant of F is given by:

$$|F| = |A|^2\sigma_1^2\sigma_2^2. \quad (20)$$

Maximize (20) is the same as maximize $|A|^2 = (\partial f/\partial w_b)^2$ with respect to d. This quantity is expressed by:

$$\left(\frac{\partial f}{\partial w_b}\right)^2 = \left(\frac{d}{cD}\right)^2 \underbrace{\left(\frac{\left(w_b\sqrt{d^2 - c^2\hat{\tau}_1^2} + cD\hat{\tau}_1\right)^2}{D^2\left(d^2 - c^2\hat{\tau}_1^2\right)} + 1\right)^{-3}}_{\xi} \quad (21)$$

Let us consider the case where the vehicle is in the broadside direction (the more convenient case for wheelbase estimation). For this case, $\hat{\tau}_1$ is very small. Since the term ξ tends to a constant when $\hat{\tau}_1$ tends to zero, it ensues that the larger the value of d, the better the estimate of w_b, but also in practice, the lower the correlation between acquired signals. This highlights that a much more precise and realistic model than (12)-(13) needs to be found to find an analytical expression of the optimal d using the CRLB. As an alternative, we explore the role of d into the BPHAT correlation function using its closed-form expression (11).



(a) $d < d_{min}$ (b) $d_{min} < d << d_{max}$

(c) $d_{min} << d < d_{max}$ (d) $d_{max} < d$

Fig. 3. Illustration of the additive effect problem as a function of the inter-sensor distance d.

B. Minimal and maximal inter-sensor distance

Because of the additive effect, due to the sum operator in Eq. (11), axles cannot be distinguished for very small values of d and phantom sources (spurious peaks) appear for very large values of d. Such an effect is depicted in Fig.3. For all plots, the acoustic scenario is the same, d being the only variable. The GCC-BPHAT function and the primary correlations are drawn in black and gray respectively. The actual TDOAs τ_1 and τ_2 and their average value τ_0 are also represented. In Fig.3(a), d is so small that it is impossible to predict the existence of the two sources. In Fig.3(b), both peaks begin to appear since d has been increased. In Fig.3(c), d has been increased again and both peaks are clearly distinct. In Fig.3(d) d has been increased again and both peaks are well distinguished but one spurious peak appears at τ_0.

As spurious peaks do not have any physical meaning here, it is always better to avoid them because of possible misinterpretations, especially when it comes to estimating the number of axles during pass-by. Consequently, the inter-sensor distance should be limited to values between a minimal distance d_{min}, above which both axles are distinct, and a maximal distance d_{max}, below which no spurious peaks appear. Inspired by Fig.3, the two peaks are distinct once $R^{bphat}_{s_1s_2}(\tau)$ is locally convex around τ_0, yielding an implicit expression of d_{min}:

$$d_{min} = \arg\min_{d>0} (g_{\tau_0} > 0) \tag{22}$$

where

$$g_{\tau_0} = \left. \frac{\partial^2 R_{s_1 s_2}^{bphat}(\tau)}{\partial \tau^2} \right|_{\tau_0} \tag{23}$$

Similarly, the condition for avoiding a central spurious peak is that $R_{s_1 s_2}^{bphat}(\tau)$ is not convex anymore around τ_0 for larger values of d. An implicit expression of d_{max} is therefore:

$$d_{max} = \arg\min_{d>d_{min}} (g_{\tau_0} < 0) \tag{24}$$

To conclude, the domain $[d_{min}, +\infty[$ defines what one can call a *range of bimodality detection*, that is, the set of inter-sensor distances for which the two peaks are observable. But in order to avoid central spurious peaks, one needs to restrict this range to $[d_{min}, d_{max}]$. We called this domain the *range of undistorted bimodality* (RUBI).

C. Range of undistorted bimodality (RUBI)

According to (22) and (24) and considering a given acoustic scenario (fixed value of D, w_b and x_0), the sign of g_{τ_0} may be expressed as a function both of the spectral properties of the BPHAT transform (B_w, f_c) and the inter-sensor distance d thanks to Eq. (11), (22) and (24). This is what Fig.4 illustrates. The vertical and horizontal axis have been specifically chosen for the sake of generalization so that spectral values are not necessarily acoustic values. This is the reason why d is normalized by the halved central wavelength $\lambda_c = c/f_c$. This plot has been generated using arbitrary geometrical parameters: w_b = 2.47 m, D = 6.3 m and x_0 = 0 m. Grey zones (respectively white zones) correspond to a negative sign (respectively positive sign) of g_{τ_0}. The six plots on the right of Fig.4 show the GCC-BPHAT at different points of the abacus (A,B,C,D,E and F).

Let us apply the BPHAT transform into the bandwidth 250-4750 Hz, *i.e.* $B_w/f_c = 1.8$. This bandwidth has been chosen empirically but any other one can be considered, depending on the application, and without undermining the theory described hereafter. In zone I, the two peaks are undetectable (point A). They begin to appear at the boundary between zone I and zone II (point B). The two peaks are clearly distinct in middle of the zone II (point C). Then, in zone III, IV and upper, secondary lobes appear around τ_0 (point D, E, F). So, in this example, the RUBI is delimited by B and D and the optimal distance d_{opt} is somewhere within this range.

In Fig.5, the same scenario as above is considered, except that the variable is now the DOA θ of the center of the vehicle (at coordinate $[x_0 + w_b/2,\ D]$) instead of the ratio B_w/f_c, the latter is fixed here to 1.8 for the whole plot. By considering the zone II, one can see that the opening angle in which bimodality is observable is more or less wide depending on d. For instance, setting $d = 5\lambda_c/2$ allows a bimodal tracking on an angle range of about 90° (±45°) as depicted by points A, B, and C. Reducing d to $3\lambda_c/2$ will reduce the observation area to nearly 70° (±35°) as depicted by points D, E and F.

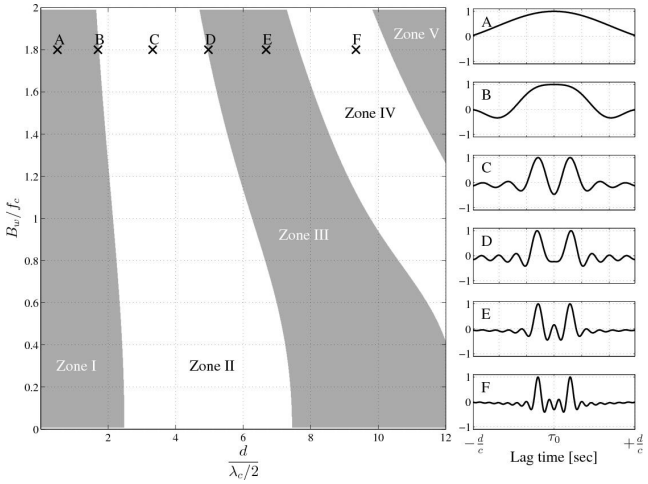


Fig. 4. Sign of g_{τ_0} as a function of the spectral properties of the BPHAT transform (B_w, f_c, λ_c) and the inter-sensor distance d. Grey (resp. white) areas correspond to a negative (resp. positive) sign.

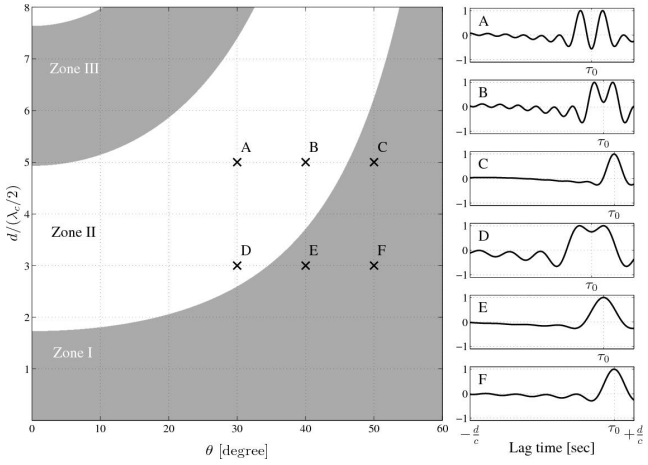


Fig. 5. Sign of g_{τ_0} as a function of d (normalized by the halved wavelength) and the vehicle direction of arrival θ in degree. The ratio B_w/f_c is set to 1.8.

D. Optimal inter-sensor distance

In passive acoustic outdoor measurements, observations are frequently corrupted because of interfering noises. Multiple vehicles may also be present in the monitoring zone, which means that the estimation of the hidden values of each vehicle (position, speed, wheelbase) had to be processed in parallel. Consequently, one good option is to rely on the Bayesian theory. In the former work [18], the successive correlation measurements are filtered by a particle filter. Particle filtering, also known as sequential Monte Carlo method, is a successful technique to recursively estimate hidden states of non-linear, non-Gaussian dynamical systems [34]. The mathematical framework of the particle filtering is not detailed here but for a good introduction, the interested reader is referred to appropriate papers such as [35], [36].

To summarize briefly, one particle is composed of a state value, *i.e.* an hypothesis, and an associated weight, *i.e.* the *probability* that this hypothesis is true regarding the observation. Recursively, each particle state is propagated by following an *a priori* dynamic model disturbed by stochastic noise and the associated weights are updated according to the observation. The more the state of a particle matches with the observation, the heavier the weight associated to this particle, and the more this particle is duplicated in favor of the lighter ones. The number of particles stays constant during all the observations.

In this study, we are looking for the optimal inter-sensor distance d_{opt} which we define as the minimal inter-sensor distance $d > d_{min}$ which guarantees the best time delay estimates by particles. This choice is driven by our objective of using a lightweight, portable, movable and small device whilst also limiting the relative Doppler effect between sensors affecting the correlation measurement. For this, the percentage error of the particle estimates w.r.t to the actual TDOA is assessed as a function of d and after a single resampling.

An illustrative example of the process is depicted in Fig.6. Let us consider a theoretical GCC-BPHAT function with two opposite TDOAs τ_1 and $-\tau_1$ and $f_c = 2500$ Hz, $B_w = 1.8 f_c$, $w_b = 2.47$ m, $D = 6.3$ m, $x_0 = w_b/2$. This function is symmetrical w.r.t 0, this is why only the positive part (in black) is represented. In (a) or (b), the shape of the correlator is typical of an appropriate inter-sensor distance, while in (c) or (d), $d > d_{max}$ therefore a central spurious speak (at $\tau = 0$) appears. At initialisation, (a) and (c), particles are uniformly distributed on the observation. After one resampling, they coalesce around the target value in (b), that is what is expected. But in (d), some particles are "attracted" by the spurious peak. In that case the convergence is not as efficient as for the previous case because both percentage error and standard deviation are higher.

In fact, Fig.6 depicts the particles distribution of one run when using two different d. The idea is to explore the statistical behavior of the particle filtering algorithm over a high number of runs (e.g. 200) and for a large set of potential d. This procedure is explained in more detail below:

1) Initialize one hundred particles uniformly on the whole states space of physically possible time delays (like in Fig.6(a) or Fig.6(c));
2) Compute the likelihood of the particles (correlation amplitude);
3) Update the particles once using, for instance, the multinomial resampling technique described in [37] section 2.1 (like in Fig.6(b) or Fig.6(d));
4) Compute and store in memory the percentage error and coefficient of variation of the particles w.r.t to the actual time delay to estimate;
5) Reiterate 1), 2) and 3) 200 times and deduce the mean percentage error and relative standard deviation of the particles distribution w.r.t to the actual time delay to estimate;
6) Reiterate 4) for each tested d.

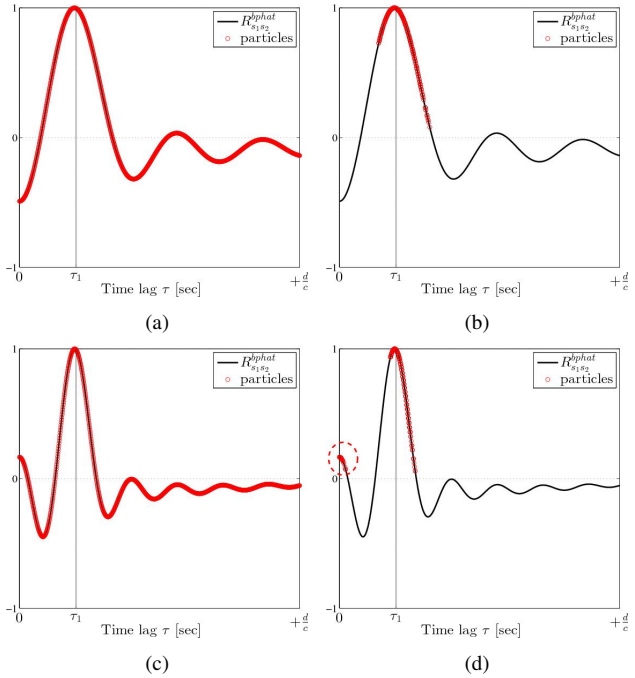


Fig. 6. Effect of a spurious peak on the particles distribution.

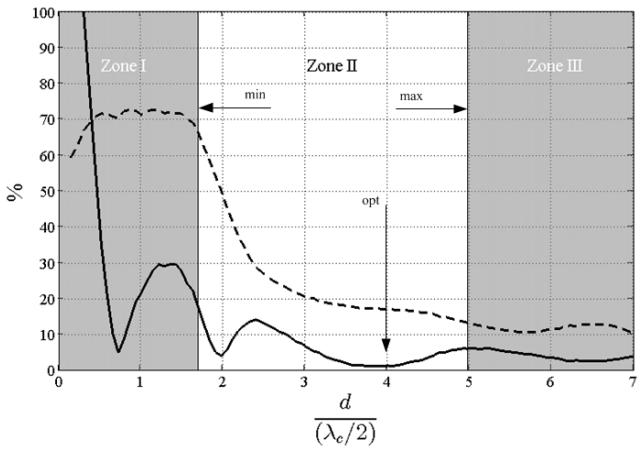


Fig. 7. Global percentage error (thick line) and coefficient of variation (dashed line) of TDOA estimation as a function of d, both expressed in %.

The result is depicted in Fig.7. As previously demonstrated, zone I should not be considered because of the non-observability of the two peaks ($d < d_{min}$). Global mean percentage error and global coefficient of variation are logically high in this area. From the beginning of the zone II (RUBI), both the accuracy and repeatability of the estimator increase. As predicted by the Fisher information matrix, the general trend is that the larger the inter-sensor distance, the better the estimate. However, with the proposed approach, one local minimum appears within the RUBI suggesting that setting $d = 2\lambda_c < d_{max}$ provides a better estimator than setting $d = d_{max}$. Hence, by integrating both the analytical model of

the correlation measure and the Monte-Carlo based tracking process in the optimization procedure correlation measure and the Monte-Carlo-based tracking process in the optimization procedure, a much more adapted design is obtained in comparison with deriving the CRLB.

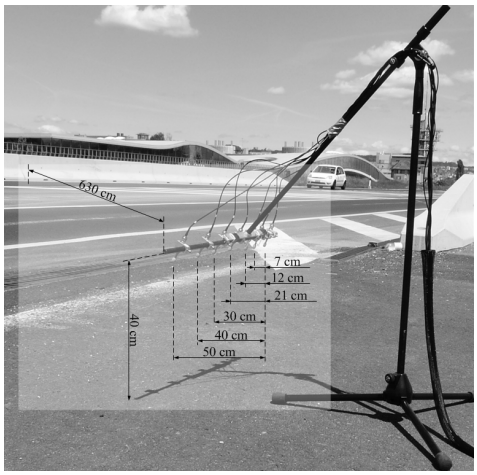
IV. Preliminary experiments

A. Experiment 1: wheelbase observation quality

A preliminary experiment has been carried out to confront the theoretical RUBI with an *in-situ* measurement. The experiment consists in acquiring the signal radiated by two moving and uncorrelated white noises using microphone pairs of different length. In order to be quite close to one realistic scenario, and also for simplicity of implementation, loudspeakers have been fixed on a car, in front of the wheels, as showed in Fig.8(a). The gain of the loudspeakers were sufficiently high for masking the own tyre/road noise of the vehicle. The wheelbase of the car was of $w_b = 2.47$ m. A linear array was disposed on the roadside at a height of 40 cm and at a distance D = 6.3 m to the loudspeakers during pass-by. The array was composed of seven microphones, allowing pairs of different apertures ranging from 7 cm to 50 cm, Fig.8(b). The vehicle speed was nearly 60 km/h during the measurement. The recording was collected on the EPFL Campus (Lat. 46°31'7.74"N, Long. 6°33'56.39"E). The location was free for reverberation but quite noisy because of a demolition site 150 meters away and a light wind. The sky was clear and the temperature was 17°C. For each pair of sensors, one BPHAT-CCTS image ($B_w/f_c = 1.8$, $f_c = 2500$ Hz) was computed. Some examples are depicted in Fig.9. Bold, respectively thin, dashed vertical lines delimit the period of time during which the vehicle is in the 60° opening angle ($-30° \leq \theta \leq +30°$), respectively 90° opening angle ($-45° \leq \theta \leq +45°$).

From Fig.4, the minimal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 1.8$, *i.e.* $d_{min} \approx 12$ cm in the present case. In Fig.9(a) and Fig.9(b), d equals 9 cm and 10 cm respectively. As expected, front and rear axles are not dissociated at all. On Fig.9(c), d equals 12 cm and one can perceive the very beginning of the separation of the two traces. This is confirmed by Fig.9(d) and Fig.9(e) in which d equals 14 cm and 18 cm respectively. From Fig.5, the minimal distance enabling the dissociation of axles over an opening angle of 60° respects the equality $d/(\lambda_c/2) \approx 2.8$, $d \approx 19$ cm . This is a rather good prediction regarding Fig.9(f) and Fig.9(g) in which d is equal to 19 cm and 21 cm respectively: the traces are well separated between the two bold dashed lines. Similarly, covering an opening angle of 90° requires d to be 31 cm. However, such an objective is actually achieved for a lower inter-sensor distance, for instance in Fig.9(h) with d equals to 28 cm. From Fig.4, the maximal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 5$, *i.e.* $d_{max} \approx 34$ cm in the present case. This is clearly demonstrated by inspecting Fig.9(i) for which d = 33 cm and Fig.9(j) for which d = 40 cm that in the first case no spurious peak appears between both traces, in opposition to the second casein which a third "phantom axle" appears between the two actual ones. Finally, from Fig.7, the optimal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 4$, *i.e.* $d_{opt} \approx 27$ cm. Indeed, one can conclude that the best contrast is achieved for d = 28 cm in this test, as shown in Fig.9(h).

(a)



(b)

Fig. 8. 1^{st} experimental setup. (a) car equipped with two loudspeakers, (b) linear array of microphones.

B. Experiment 2: wheelbase length estimation

A second measurement was carried out on the Route Cantonale of Ecublens, near the EPFL campus (Latitude 46°31'0.28"N, Longitude 6°33'50.41"E). A two-element microphone array was set up on the roadside at a height of 84 cm and at an average distance of D = 2.5 m to the vehicles closest wheels. The optimal inter-sensor distance provided by the presented method is d_{opt} = 20.4 cm thus we opted for d = 20 cm. The scene was continuously filmed by two cameras, one placed on the road side near the radar to get a view of the sides of all the vehicles and another placed on the balcony of a nearby building to get a more global view of the scene. Both devices produced video at 30 frames per second. Fig.10 depicts the views provided by both cameras and the location of the microphone array. Only the right-hand traffic lane is considered in this experiment, namely the lane where a black vehicle is present on these pictures. Audio and video signals are synchronized off-line. An home-made detection algorithm through successive image differences in the square of Fig.10 returned the time of apparition of each new vehicle in this

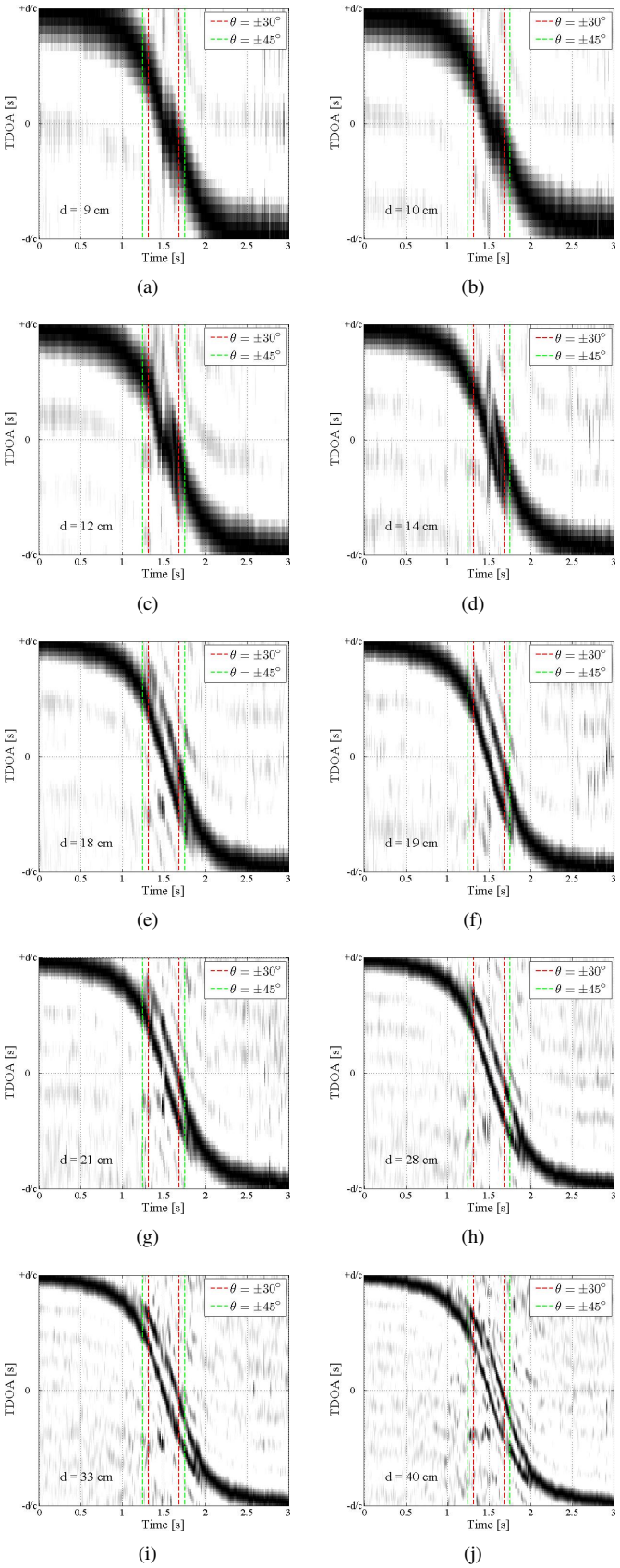


Fig. 9. GCC-BPHAT time series of the same vehicle pass-by using different array apertures.

lane. The recording takes 240 seconds. The sampling rate is f_s = 51.2 kHz. During this time, 24 vehicles were detected. The brand and model of each vehicle was identified so that their actual wheelbase length is assumed to be known.

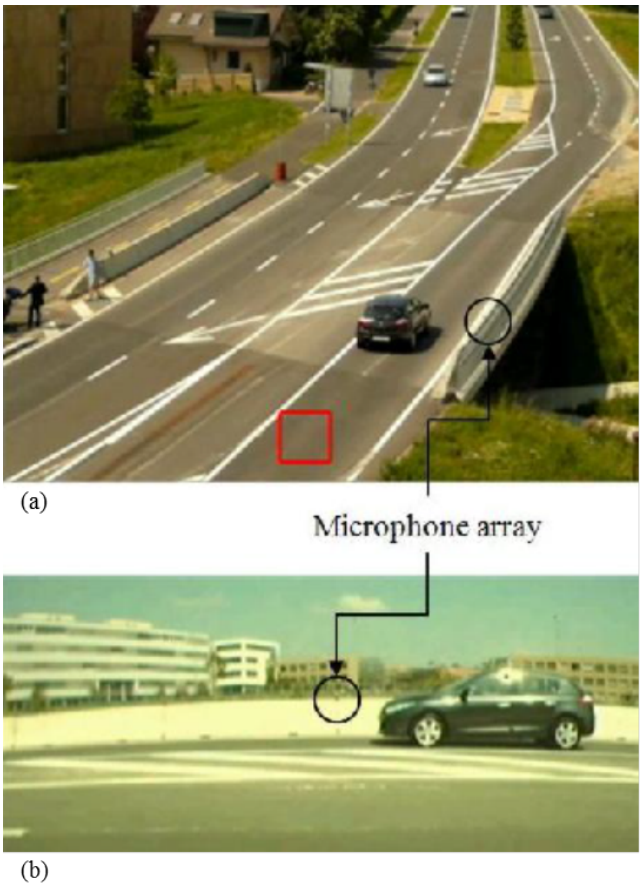


Fig. 10. 2^{nd} experimental setup. The location of the microphone array is highlighted by a black circle. In these pictures, the vehicle of interest is viewed from above (a) and the side (b) using two cameras.

The whole CCTS is constituted of successive GCC-BPHAT function applied on short audio frames of length N_s = 2048 samples (40 ms), 75% overlapped (30 ms). For each passage, the speed and wheelbase length are estimated using the bimodal particle filter proposed in [18] with 5000 particles. Performances are averaged over 200 runs.

Results are depicted in Fig.11. The acoustic estimates (circles) are compared to the actual ones (crosses) and their absolute differences are represented by a bar chart below. The *a priori* wheelbase length is 2.25 m has also been represented by a horizontal dashed line. For clarity, actual wheelbase lengths have been sorted in ascending order. We should mention that the *a priori* wheelbase length value has been arbitrary chosen but permits also to show the robustness of the method as this *a priori* is quite far from actual values.

Despite an *a priori* wheelbase length far below the actual ones, the estimates are pretty good for wheelbase lengths varying between 2.4 m and 2.8 m. When the axles are poorly

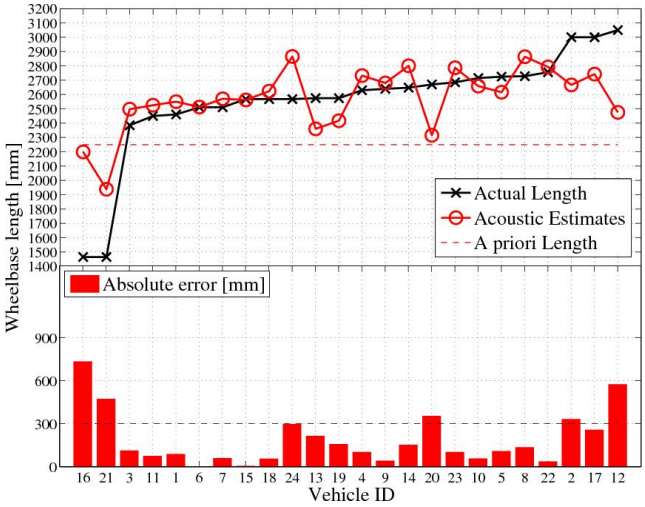


Fig. 11. Confrontation between actual and acoustic wheelbase estimates as a function of the vehicle ID when using two microphones.

observed, the final result tends to be close to the *a priori* value. This is the case for vehicles 16 and 21 which are in fact motorbikes. For such vehicles, the tyre/road noise is dominated by the exhaust noise so that only one trace appears in the CCTS, making the wheelbase length estimation impossible. The estimates are also poor for cars 2, 12 and 17 since their wheelbases are too distant from the *a priori* value. After excluding the two motorbikes from the database, as they are considered out of context, we obtained an error less than or equal to 30 cm for 19 out of 22 cases. It is worth noting that these results not only depend of the quality of the observation (which has been optimized according to the methodology described in this paper) but also on the detection strategy (defining the initial conditions of the target) and particle filter parameters (number of particles, noise covariance matrix, dynamical model, likelihood model etc.). Therefore much better scores can be expected by also optimizing these two other aspects.

V. Conclusion

In this paper, we confirmed that a well designed pair of microphones, placed on the roadside, enables the wheelbase length of two-axle road vehicles to be estimated during pass-by. The presented work related to the way of optimizing the inter-sensor distance so as to to improve the cross-correlation-based observations of both axles trajectories. The wheelbase length estimation problem is primarily a time-delay estimation problem. Due to the additive effect occurring in the cross-correlation in presence of more than one source, the model (12)-(13) appears to be too simplistic compared to simulated and experimental results since the variance and error of time-delay estimates do not evolve quadratically with the inter-sensor distance. A heuristic methodology of design has therefore been proposed consisting in 1) expressing the closed-form expression of the observation, 2) defining a range within which the inter-sensor distance must be contained, 3) filtering the modeled observation with a sequential Monte-Carlo method for each inter-sensor distance within this range and 4) looking at which candidates yield the most accurate and repeatable time-delay estimates.

Experimental measurements have been designed to confirm the difficulties and to validate the proposed approach. In particular, a preliminary test of wheelbase length estimation on 22 unknown vehicles passing-by has been carried out, after following the proposed array design methodology. An error of less than 30 cm was obtained in 86% of the cases, *i.e.* less than the size of a wheel, which is rather promising since only two sensors have been used.

The acoustic-based wheelbase estimation is still in its infancy, but even better results can be expected over the coming years. Further research is likely to improve the observation quality using an array with more than two sensors so to exploit the redundant information between sensor pairs.

Appendix A
Closed-form expression of the GCC-BPHAT function in the single source case

Without noise and under free field conditions, the signal acquired by one sensor is a delayed version of the signal acquired by the other sensor, such that:

$$x_2(t) = x_1(t + \tau_1). \tag{25}$$

Eq. (25) may be translated to the frequency domain by:

$$X_2(f) = X_1(f)e^{+2j\pi f\tau_1}, \tag{26}$$

where X_i(f) and x_i(t) are related by the Fourier and inverse Fourier transform according to the conventions:

$$X_i(f) = \int_{-\infty}^{+\infty} x_i(t)e^{-2j\pi ft}dt, \tag{27}$$

$$x_i(t) = \int_{-\infty}^{+\infty} X_i(f)e^{+2j\pi ft}df. \tag{28}$$

Substituting (26) into the expression of the GCC (5) with $\psi_g(f)$ is replaced by the BPHAT weighting $\psi_{bphat}(f)$ (8) gives:

$$R_{s_1s_2}(\tau) = \int_{-\infty}^{+\infty} \frac{X_1X_1^*}{|X_1X_1^*|}e^{2j\pi f(\tau-\tau_1)}df, \tag{29}$$

$$= \int_{-f^+}^{-f^-} e^{2j\pi f(\tau-\tau_1)}df + \int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_1)}df,$$

$$= 2\mathcal{R}\left[\int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_0)}df\right], \tag{30}$$

where $\mathcal{R}[.]$ is the real part operator. Futhermore:

$$\int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_0)}df = \frac{e^{2j\pi f^+(\tau-\tau_0)} - e^{2j\pi f^-(\tau-\tau_0)}}{2j\pi(\tau-\tau_0)},$$

$$= \frac{e^{j\pi\gamma}sin(\pi\gamma)}{\pi(\tau-\tau_0)}, \tag{31}$$

where $\gamma = (f^+ + f^-)(\tau - \tau_0)$. Replacing $f^+ + f^-$ by $2f_c$ and $f^+ - f^-$ by B_w yields the expression (9).

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their suggestions of improvements as well as Patrick Roe for his proofreading and Cédric Monchâtre for his help in the experimental measurements.

REFERENCES

[1] H. C. Choe, R. E. Karlsen, G. R. Gerhart, and T. J. Meitzler, "Wavelet-based ground vehicle recognition using acoustic signals," *Proceedings of SPIE*, vol. 2762, pp. 434–445, 1996.

[2] S. Chen, Z. Sun, and B. Bridge, "Automatic traffic monitoring by intelligent sound detection," in *Proceedings of the IEEE Conference on Intelligent Transportation System (ITSC)*, dec. 1997, pp. 171–176.

[3] J. F. Forren and D. Jaarsma, "Traffic monitoring by tire noise," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC)*, 1997, pp. 177–182.

[4] E. Brockmann, B. Kwan, and L. Tung, "Audio detection of moving vehicles," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics.*, vol. 4, oct. 1997, pp. 3817–3821.

[5] J. P. Kuhn, B. C. Bui, and G. J. Pieper, "Acoustic sensor system for vehicle detection and multi-lane highway monitoring," International Patent 5 798 983, aug, 1998.

[6] S. Chen, Z. Sun, and B. Bridge, "Traffic monitoring using digital sound field mapping," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 6, pp. 1582–1589, nov. 2001.

[7] K. Kodera, A. Itai, and H. Yasukawa, "Approaching vehicle detection using linear microphone array," in *Proceedings of International Symposium on Information Theory and Its Applications (ISITA)*, dec. 2008, pp. 1–6.

[8] C. Kwak, M. Kim, K. Kim, S. Hong, and K. Kim, "Robust in-situ vehicle detection algorithm with acoustic transition bandpass filter," feb. 2009.

[9] N. Shimada, A. Itai, and H. Yasukawa, "A study on using linear microphone array-based acoustic sensing to detect approaching vehicles," in *Proceedings of International Symposium on Communications and Information Technologies (ISCIT 2010)*, oct. 2010, pp. 182–186.

[10] B. Barbagli, L. Bencini, I. Magrini, G. Manes, and A. Manes, "A real-time traffic monitoring based on wireless sensor network technologies," *Proceedings of the 7^{th} International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 820–825, jul. 2011.

[11] J. Towers and Y. Chan, "Passive localization of an emitting source by parametric means," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, apr. 1990, pp. 2791–2794.

[12] C. Couvreur and Y. Bresler, "Doppler-based motion estimation for wide-band sources from single passive sensor measurements," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, apr. 1997, pp. 3537–3540.

[13] F. Pérez-González, R. López-Valcarce, and C. Mosquera, "Road vehicle speed estimation from a two-microphone array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, apr. 2002.

[14] R. López-Valcarce, C. Mosquera, and F. Pérez-González, "Estimation of road vehicles speed using two omnidirectional microphones: a maximum likelihood approach," *EURASIP Journal on Applied Signal Processing*, vol. 8, pp. 1059–1077, 2004.

[15] O. Duffner, N. O'Connor, N. Murphy, A. Smeanton, and S. Marlow, "Road traffic monitoring using a two-microphone array," in *Audio Engineering Society, Convention 118*, may 2005, p. 6355.

[16] V. Cevher, R. Chellappa, and J. McClellan, "Vehicle speed estimation using acoustic wave patterns," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 30–47, jan. 2009.

[17] A. Can, L. Dekoninck, M. Rademaker, T. V. Renterghem, B. D. Baets, and D. Botteldooren, "Noise measurements as proxies for traffic parameters in monitoring networks," *Science of The Total Environment*, vol. 410-411, pp. 198–204, 2011.

[18] P. Marmaroli, J.-M. Odobez, X. Falourd, and H. Lissek, "A bimodal sound source model for vehicle tracking in traffic monitoring," in *Proceedings of the 19^{th} European Signal Processing Conference (EUSIPCO)*, 2011, pp. 1327–1331.

[19] B. Yang and J. Scheuing, "Cramer-rao bound and optimum sensor array for source localization from time differences of arrival," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2005, pp. 961–964.

[20] J.-S. Hu, C.-M. Tsai, C.-Y. Chan, and Y.-J. Chang, "Geometrical arrangement of microphone array for accuracy enhancement in sound source localization," in *Proceedings of the 8^{th} Asian Control Conference (ASCC)*, may 2011, pp. 299–304.

[21] U. Sandberg, "Tyre/road noise - myths and realities," in *Proceedings of the 2011 International Congress and Exhibition on Noise Control Engineering*, 2001.

[22] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I*. Wiley-Interscience, 2001.

[23] 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.

[24] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer US, 2004, pp. 197–227.

[25] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?" in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, mar. 2008, pp. 2565–2568.

[26] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[27] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, nov. 2003.

[28] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 2, oct. 2003, pp. 1228–1233.

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

[30] Y. Gao, M. Brennan, P. Joseph, J. Muggleton, and O. Hunaidi, "A model of the correlation function of leak noise in buried plastic pipes," *Journal of Sound and Vibration*, vol. 277, Issues 1-2, pp. 133–148, oct. 2004.

[31] Y. Gao, M. Brennan, and P. Joseph, "A comparison of time delay estimators for the detection of leak noise signals in plastic water distribution pipes," *Journal of Sound and Vibration*, vol. 292, Issues 3-5, pp. 552–570, 2006.

[32] S. M.Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall PTR, 2010.

[33] R. Mehra, "Optimal input signals for parameter estimation in dynamic systems - survey and new results," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 753–768, dec. 1974.

[34] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[35] P. Djuric, J. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, sep 2003.

[36] J. Candy, "Bootstrap particle filtering," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 73–85, jul. 2007.

[37] R. Douc, O. Cappé, and E. Mou, "Comparison of resampling schemes for particle filtering," in *Proceedings of the 4^{th} International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2005.

Patrick Marmaroli was born in Saint-Julien-en-Genevois, France, in 1984. He received a M.Sc. degree in signal processing and trajectography from Sud-Toulon Var University, France, in 2008. The same year, he enrolled as a PhD student at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His current research interests include acoustic array processing for denoising, localization and multi target tracking.

Mikael Carmona was born in Grenoble, France, in 1984. He graduated as an engineer (2007) and Ph.D. (2011) in signal processing from the Institut National Polytechnique de Grenoble. He also graduated in mathematics (Agrégation) in July 2007 at the Universit Joseph Fourier de Grenoble. Since 2011, he has been engineer at the Commissariat l'Energie Atomique et aux Energies Alternatives. He is working on the reconstruction of geometrical and physical parameters from inertial sensors (accelerometers, magnetometers, gyrometers) and from seismic/acoustic sensors.

Jean-Marc Odobez is graduated from the Ecole Nationale Supérieure de Télécommunications de Bretagne (ENSTBr) in 1990, and received his Ph.D degree in Signal Processing and Télécommunication from Rennes University, France in 1994. He performed his dissertation research at IRISA/INRIA Rennes on dynamic scene analysis using statistical models. He then spent one year as a post-doctoral fellow at the GRASP laboratory, University of Pennsylvania, USA, working on visually guided robotic navigation problems. From 1996 until September 2001, he was associate professor in computer science at the Université du Maine, France. He his now a senior researcher at both the IDIAP Research Institute and Ecole Polytechnique Fédéral de Lausanne (EPFL), Switzerland. He has worked for several years on the development of computer vision, machine learning and statistical models for image representation and segmentation, object recognition, tracking, human activity recognition and multimedia content analysis. He is or has been Principal Investigator for three European research projects and four swiss projects, and co-PI of three others. He is author and coauthor of more than 100 papers in international journals and conferences, and is associate editor of the Machine Vision and Application journal. He is holding 2 patents on video motion analysis. He is the co-founder of the Swiss Klewel SA company active in the intelligent capture, indexing, and webcasting of multimedia conference and seminar events.

Xavier Falourd was born in La Roche Sur Yon, France, in 1974. He graduated as a Bachelor (Master) in fundamental mechanics from Université de Poitiers, France, in 1998, and a Research Master (DEA) in mechanics and acoustics from Ecole Centrale de Nantes, France in 1999. In the Laboratoire d'Electromagnétisme et d'Acoustique (LEMA) he worked on underwater acoustics to study large scale hydrodynamics of the lake Geneva using acoustic tomography and he received a PhD degree from Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2004, with a speciality in acoustics. He is currently a Postdoctoral Research Associate in the Acoustic Group of the LEMA at EPFL, working on numerous applied fields of acoustics and array processing.

Hervé Lissek was born in Strasbourg, France, in 1974. He graduated in fundamental physics from Université Paris XI, Orsay, France, in 1998, and received the Ph.D. degree from Université du Maine, Le Mans, France, in July 2002, with a speciality in acoustics. From 2003 to 2005, he was a Research Assistant at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, with a specialization in electroacoustics and active noise control. Since 2006, he has been heading the Acoustic Group of the Laboratoire d'Electromagnétisme et d'Acoustique at EPFL, working on numerous applied fields of electroacoustics and audio engineering.