Time Difference of Arrival Estimation based on Binary Frequency Mask for Sound Source Localization on Mobile Robots

François Grondin and François Michaud

***Abstract*— Localization of sound sources in adverse environments is an important challenge in robot audition. The target sound source is often corrupted by coherent broadband noise, which introduces localization ambiguities as noise is often mistaken as the target source. To discriminate the time difference of arrival (TDOA) parameters of the target source and noise, this paper presents a binary mask for weighted generalized cross-correlation with phase transform (GCC-PHAT). Simulation and experiments on a mobile robot suggest that the proposed technique improves TDOA discrimination. It also brings the additional benefit of modulating the computing load requirement according to voice activity.**

***Index Terms*— TDOA, GCC-PHAT, binary mask, sound source localization, robot audition**

I. Introduction

Robot audition provides important cues for mobile robots to interact with their environment. Computational Auditory Scene Analysis (CASA) consists in, amongst other, localizing and separating a mixture of sound sources. Speech recognition and speaker identification are then performed on these separated sound sources [1], [2], [3]. Localization is usually the first step in the robot audition processing sequence, as it allows the robot to focus its attention toward a specific direction and use the localization information to enhance sound source separation.

Delay-and-sum beamformer is a popular technique to perform sound source localization with a microphone array [4]. This method relies on time difference of arrival (TDOA) estimation between each pair of microphones. Generalized Cross-Correlation with Phase Transform (GCC-PHAT) is usually used to perform TDOA estimation. Its simplicity and robustness to high reverberant environment makes it suitable for robot audition applications. Although this method is robust to reverberation, it is sensitive to broadband additive noise. To reduce the noise contribution, a non-recursive frequency mask is proposed by Valin et al. [5]. This mask is computed according to the instantaneous signal-to-noise ratio (SNR) and has unbounded maximum value, which makes the result highly dependent on the volume of the target source signal. For this reason, the recursive mask was introduced with bounded values and has been used extensively in the ManyEars system [6], [7]. However, the system remains sensitive to broadband directional (coherent) noise and requires continuous processing even during silence periods.

To improve robustness to noise, the Multiple Signal Classification based on Standard Eigenvalue Decomposition (SEVD-MUSIC) approach, initially used for narrowband signals [8], has been adapted for broadband sound source localization [9]. The idea is to decompose the covariance matrix obtained from the spectral observations at each microphone in both noise and noisy signal subspaces. The direction of the sound source is obtained by finding one or many direction of arrival vectors orthogonal to the noise subspace. This method is efficient as long as the noise is less powerful than the signals to be localized. To deal with this issue, Multiple Signal Classification based on Generalized Eigenvalue Decomposition (GEVD-MUSIC) is a possible alternative [10]. Although SEVD-MUSIC and GEVD-MUSIC improve significantly the robustness to noise, they have two limitations: the performance drops under significant reverberation, and the eigenvalue decomposition leads to high computational load. As a solution, Multiple Signal Classification based on Generalized Singular Value Decomposition (GSVD-MUSIC) is introduced to reduce computational load [11]. This technique also improves the localization accuracy as the eigenvectors are mutually orthogonal, but still remains sensitive to highly reverberant environments.

Specific geometries (e.g., line, circular, and spherical microphone dispositions) have also been studied to improve performance and reduce computational load [12], [13], [14]. For mobile robots, these configurations are not ideal due to the physical constraints introduced by the shape of the robot.

Beamforming based on weighted GCC-PHAT is an appealing method to perform sound source localization as it is robust to reverberation, can be used with microphone arrays with arbitrary shapes, has low-complexity, and is robust to additive noise when the GCC-PHAT weighting mask is optimal. This paper introduces a new binary frequency mask used to enhance the performance of weighted GCC-PHAT under broadband coherent noise, and that can modulate computational load with voice activity. This paper is organized as follows. Section II describes the proposed method, followed by section III with results obtained. Section IV concludes the paper with ideas for future work.

II. Weighted GCC-PHAT with Binary Frequency Mask

In this section, we first define a model to pinpoint how coherent noise influences TDOA estimation results, and then describe the weighted GCC-PHAT method and the new

This work was supported by the Fonds de recherche du Québec - Nature et technologies (FRQNT).

F. Grondin and F. Michaud are with the Department of Electrical Engineering and Computer Engineering, 3000 boul. de l'Université, Sherbrooke, Québec (Canada) J1K 2R1, {Francois.Grondin2,Francois.Michaud}@USherbrooke.ca

binary mask that improves robustness to noise. We then discuss the computational load introduced by beamforming with a matrix of microphones and how the proposed method can improve computational efficiency.

A. Model

For this scenario, it is assumed that the target sound source is corrupted by coherent and incoherent noises, and is captured in a reverberant environment. Coherent noise originates from a specific direction, which implies that its observations on multiple microphones are correlated. Incoherent noise is diffuse, and observations on multiple microphones are uncorrelated. The target sound source is represented by the signal $s[n]$, and $h_m[n]$ is the room impulse response (RIR) between this source and each microphone m. The variable n stands for the sample index in time. Coherent noise is modelized as a noise source $c[n]$ convolved with a RIR $g_m[n]$, while incoherent noise is represented by a single additive term $b_m[n]$. The captured signal $x_m[n]$ at each microphone m is expressed in (1), with $*$ standing for the convolution operator.

$$x_m[n] = h_m[n] * s[n] + g_m[n] * c[n] + b_m[n] \tag{1}$$

The observed signals in the frequency domain are shown in (2), with ω and j standing for the normalized frequency in radians and the complex number $\sqrt{-1}$, respectively. This is an approximation since the target sound source is only stationnary during intervals of time shorter than the duration of the RIR.

$$X_m(e^{j\omega}) = H_m(e^{j\omega})S(e^{j\omega}) + G_m(e^{j\omega})C(e^{j\omega}) + B_m(e^{j\omega}) \tag{2}$$

The cross-correlation between the signals of two microphones ($p \neq q$) in (3) carries information about the position of the target sound source.

$$\begin{aligned} \mathrm{E}\{X_p(e^{j\omega})X_q(e^{j\omega})^*\} \quad = \quad & H_p(e^{j\omega})H_q(e^{j\omega})^*|S(e^{j\omega})|^2 + \\ & G_p(e^{j\omega})G_q(e^{j\omega})^*|C(e^{j\omega})|^2 \end{aligned} \tag{3}$$

The term $\mathrm{E}\{...\}$ stands for the expected value operator. While the incoherent noise term $\mathrm{E}\{B_p(e^{j\omega})B_q(e^{j\omega})^*\}$ vanishes to zero for $p \neq q$, the coherent noise is preserved and may overshadow the target sound source term if the power of the coherent noise $|C(e^{j\omega})|^2$ is greater than the target sound source power $|S(e^{j\omega})|^2$. The coherent noise may generate a persistent TDOA value that is mistaken as the target source TDOA value.

B. Weighted GCC-PHAT

TDOA estimation is performed here with a weighted GCC-PHAT, which uses a short-time Fourier transform (STFT). The STFT is computed using a Fast Fourier Transform (FFT) for each frame of N samples, as given by (4). A Hann window $w[n]$ is used and the frames are spaced by a hop size of ΔN samples. The variables l and k stand for the frame and the frequency bin indexes, respectively.

$$X_m^l[k] = \sum_{n=0}^{N-1} w[n]x_m[l\Delta N + n]e^{-j2\pi kn/N} \tag{4}$$

The weighted GCC-PHAT is expressed in (5). The variables $X_p^l[k]$ and $X_q^l[k]$ stand for the STFT coefficients of channels p and q, respectively. The masks $\zeta_p^l[k]$ and $\zeta_q^l[k]$ emphasize on the target frequency bins to reduce noise contribution. The variable ϵ is added to avoid overflow when the expression $|X_p^l[k]||X_q^l[k]|$ goes to zero. The operator $(...)^*$ stands for the complex conjugate. The weighted GCC-PHAT is efficiently computed with an Inverse Fast Fourier Transform (IFFT).

$$R_{p,q}^l[n] = \frac{1}{N}\sum_{k=0}^{N-1} \frac{\zeta_p^l[k]X_p^l[k]\zeta_q^l[k]X_q^l[k]^*}{|X_p^l[k]||X_q^l[k]| + \epsilon} e^{j2\pi kn/N} \tag{5}$$

The GCC-PHAT method is efficient against reverberation as long as the target sound source is broadband and dominates most frequency bins. The frequency masks $\zeta_p^l[k]$ and $\zeta_q^l[k]$ are designed to enhance these bins. The unwrapped estimated TDOA value $\hat{\tau}_{p,q}^l$ corresponds to the sample index n with the maximum value of the weighted GCC-PHAT result, as expressed by (6).

$$\hat{\tau}_{p,q}^l = \arg\max_n \left(R_{p,q}^l[n]\right) \tag{6}$$

The variable $\hat{\tau}_{p,q}^l$ lies in the range $[0, N-1]$. It is more convenient to express TDOA as a wrapped value that lies within the range $[-N/2, (N/2-1)]$, and the estimated TDOA is therefore obtained by (7). The expression (...) mod (...) stands for the modulo operation.

$$\tau_{p,q}^l = \left\{\left(\hat{\tau}_{p,q}^l + \frac{N}{2}\right) \bmod N\right\} - \frac{N}{2} \tag{7}$$

The corresponding energy is obtained by (8). Equation (9) shows that energy is used to discriminate true detection when it is greater or equal to the threshold E_0 ($e^l = 1$), and false detection ($e^l = 0$) otherwise.

$$E_{p,q}^l = R_{p,q}^l[\hat{\tau}_{p,q}^l] \tag{8}$$

$$e^l(E_0) = \begin{cases} 0 & E_{p,q}^l < E_0 \\ 1 & E_{p,q}^l \geq E_0 \end{cases} \tag{9}$$

Two dominant TDOA values are observed when a target source is corrupted by broadband coherent noise. We would like to extract the target source TDOA (τ_t) and ignore the coherent noise TDOA (τ_c). A TDOA $\tau_{p,q}^l$ is assigned to the target source ($t^l = 1$) when the absolute value of the difference with τ_t is less or equal to a constant $\Delta\tau$, as given by (10). Equation (11) shows that the TDOA $\tau_{p,q}^l$ is classified as a coherent noise TDOA ($c^l = 1$) in the same way.

$$t^l = \begin{cases} 0 & |\tau_{p,q}^l - \tau_t| > \Delta\tau \\ 1 & |\tau_{p,q}^l - \tau_t| \leq \Delta\tau \end{cases} \tag{10}$$

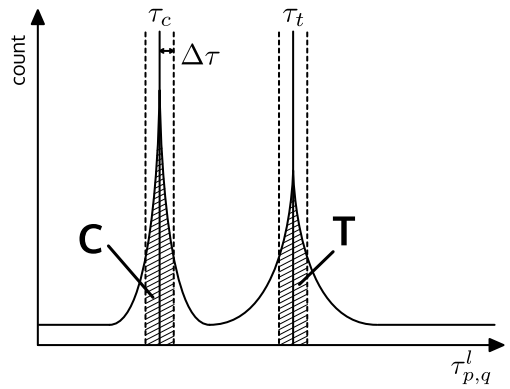


Fig. 1: $\tau^l_{p,q}$ distribution for a target source under broadband coherent noise.

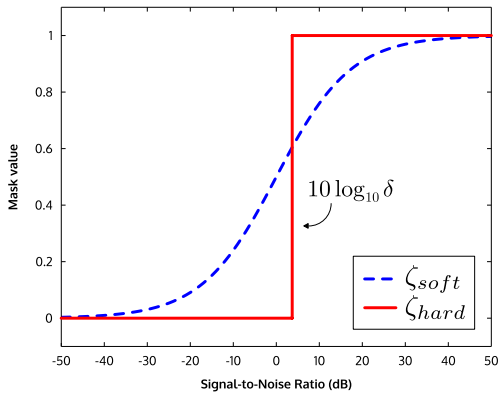


Fig. 2: Mask value according to signal-to-noise ratio.

$$c^l = \begin{cases} 0 & |\tau^l_{p,q} - \tau_c| > \Delta\tau \\ 1 & |\tau^l_{p,q} - \tau_c| \leq \Delta\tau \end{cases} \tag{11}$$

The objective here is to maximize T, the sum of t^l introduced in (12), while minimizing C, the sum of c^l expressed by (13). These sums correspond to the areas under the curve C and T shown in Fig. 1. To measure the capacity of the system to discriminate target source TDOA values from coherent noise TDOA values, the objective is to maximize the ratio $r_{classification}$ given by (14).

$$T(E_0) = \sum_{l=0}^{L-1} e^l(E_0)t^l \tag{12}$$

$$C(E_0) = \sum_{l=0}^{L-1} e^l(E_0)c^l \tag{13}$$

$$r_{classification}(E_0) = \frac{T(E_0)}{T(E_0) + C(E_0)} \tag{14}$$

It is possible to bring the latter ratio to 1 by keeping only the few TDOA values with the highest energy. Although the classification is accurate, the ratio of TDOA values used over the total number of frames is low, and this makes source localization difficult over time. We therefore introduce in (15) a second metric, $r_{detection}$, which is the ratio of $T(E_0)$ over the total number of frames L.

$$r_{detection}(E_0) = \frac{T(E_0)}{L} \tag{15}$$

C. Masks

Valin et al. [6] use a soft mask which values range from 0 to 1 according to the estimated SNR. The expression $\xi^l_m[k]$ is an estimate of the *a priori* SNR at the m^{th} microphone, obtained with the method proposed by Ephraim and Malah [15]. The expression $\sigma^2_m[k]$ stands for the stationary noise estimate obtained with the Minima-Controlled Recursive Average (MCRA) method [16]. The soft mask $(\zeta_{soft})^l_m[k]$ and the *a priori* SNR are computed recursively using (16) and (17). The parameter α_D stands for the adaptation rate.

$$(\zeta_{soft})^l_m[k] = \frac{\xi^l_m[k]}{\xi^l_m[k] + 1} \tag{16}$$

$$\xi^l_m[k] = \frac{(1-\alpha_D)(\zeta_{soft})^{l-1}_m[k]^2 |X^{l-1}_m[k]|^2 + \alpha_D |X^l_m[k]|^2}{\sigma^2_m[k]} \tag{17}$$

With $(\zeta_{soft})^l_m[k]$, broadband coherent noise may leak from each bin if the mask value is nonzero, and generate an undesired dominant peak in the TDOA estimation result. To solve this issue, we propose a new mask that behaves as a binary mask, which we call hard mask $(\zeta_{hard})^l_m[k]$, that takes only discrete values of 0 or 1. Each $(\zeta_{hard})^l_m[k]$ is set to 1 when the SNR $10\log_{10}\left(|X^l_m[k]|^2/\sigma^2_m[k]\right)$ exceeds the threshold $10\log_{10}\delta$, and to 0 otherwise, as expressed by (18).

$$(\zeta_{hard})^l_m[k] = \begin{cases} 0 & 10\log_{10}\left(\dfrac{|X^l_m[k]|^2}{\sigma^2_m[k]}\right) \leq 10\log_{10}\delta \\ 1 & 10\log_{10}\left(\dfrac{|X^l_m[k]|^2}{\sigma^2_m[k]}\right) > 10\log_{10}\delta \end{cases} \tag{18}$$

The mask values are plotted in Fig. 2 as a function of the estimated SNR. This figure illustrates how the hard mask entirely removes undesirable noise below the given threshold $10\log_{10}\delta$ (in dB).

D. Computation Load

It is also relevant to consider the computation load introduced by the TDOA estimation process for a matrix of microphones. For M microphones, there are $M(M-1)/2$ pairs, and the same number of GCC-PHAT computations. At each frame, M FFTs and $M(M-1)/2$ IFFTs are computed. Assuming both transforms involve the same amount of computer cycles, the computation load corresponds to $M + M(M-1)/2$ FFTs, which simplifies to $(M^2+M)/2$ FFTs, and a complexity of $O(M^2)$. As the number of microphones in the matrix increases, the complexity becomes significant and this eventually overloads the robot onboard processor. The SNR at each microphone may vary according

TABLE I: Positions for the simulations

Emitter / Receiver	x (m)	y (m)	z (m)
Microphone 1	-0.30	-0.30	+0.50
Microphone 2	-0.20	-0.20	+0.25
Speech source	+1.00	+0.50	+0.25
Noise source	-1.00	-0.25	+0.50

to the matrix configuration and the source position, and the microphones with noisy signals may be neglected in order to reduce the amount of computations. Moreover, during silence periods, the TDOA estimation process may be paused to decrease computational load and free computing resources for other purposes. As mentioned in Section II-B, the weighted GCC-PHAT method is efficient in a reverberant environment as long as the target sound source is broadband and dominates most frequency bins. Therefore, a minimum number of bins must be excited by the target source to obtain a relevant result from the weighted GCC-PHAT transform. The number of significant bins (θ_m^l) is obtained from the hard mask according to (19).

$$\theta_m^l = \sum_{k=0}^{N/2} (\zeta_{hard})_m^l[k] \tag{19}$$

Equation (20), which now replaces (8) when the mask ζ_{hard} is used, shows that the GCC-PHAT only has to be computed for the pair of microphones p and q when θ_m^l is equal or greater to the threshold θ_{min}. When θ_m^l is smaller, the system does not compute the GCC-PHAT with (5), and discards the TDOA by setting the energy level to 0.

$$E_{p,q}^l = \begin{cases} R_{p,q}^l[\hat{\tau}_{p,q}^l] & \theta_p^l \geq \theta_{min}, \theta_q^l \geq \theta_{min} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

III. Results

First, we validated the proposed method with simulations that cover a wide range of reverberation levels and SNRs, to then conduct experiments on a mobile robot.

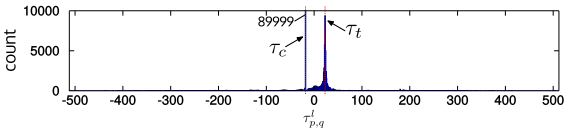
1) Simulations: We tested the weighted GCC-PHAT system with different masks to characterize the performances of each mask. The RIRs for reverberation time parameters (RT60) of 0, 250, 500, 750, and 1000 msec are generated for a room of dimensions 10 m × 6 m × 3 m with the Allen and Berkley image method [17]. Two microphones capture the simulated signals, and their positions are given in Table I. The target source comes from male and female speech segments separated by silence periods of 1 sec, and convolved with the associated RIRs. A white gaussian noise is generated and convolved with the corresponding RIRs to simulate the directional (coherent) noise. The positions of the target and noise sources are given in Table I. The z axis points to the ceiling of the room and the origin of this Cartesian system is positioned in the center of the room.

Parameters used in the simulations are given in Table II. The adaptation rate α_D used to compute ζ_{soft} matches the value proposed in [6]. The threshold δ introduced with

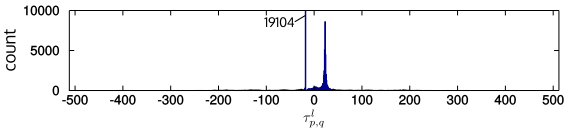
TABLE II: Parameters

Parameters	f_s	ΔN	N	ϵ	α_D	δ
Values	48000	512	1024	1E-20	0.1	5.0

(a) $E_0 = 0.001$



(b) $E_0 = 0.019$



(c) $E_0 = 0.039$

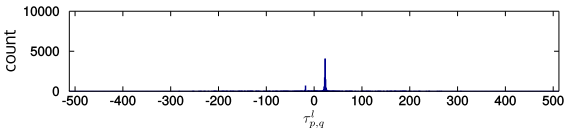


Fig. 3: $\tau_{p,q}^l$ distribution according to E_0 with ζ_{soft}.

ζ_{hard} needs to be large enough to discriminate between noise and the target sound source, and small enough to include a minimum number of bins excited by the target sound source. The high sample rate (f_s) allows the acquisition of speech high frequency components. N is the same as in [6] and the hop size ΔN ensures an overlap of 50%. The parameter ϵ prevents overflow and is small enough to preserve the precision of (5).

We present the detailed results of a specific case (when RT60 = 500 msec, SNR = 10 dB and $\theta_{min} = 0$) to give some insights about the TDOA distributions and to illustrate how performance is measured. Figure 3 shows the histograms of $\tau_{p,q}^l$ for TDOA values that satisfy condition in (9) for multiple values of E_0, when soft masks are used. In Fig. 3a and 3b, we observe that the distribution of $\tau_{p,q}^l$ peaks at τ_c, which shows that the coherent noise term dominates significantly. As the energy level E_0 increases in Fig. 3c, the undesired peak vanishes and τ_t then becomes the dominant peak.

Figure 4 shows the histograms of $\tau_{p,q}^l$ for multiple values of E_0, when hard masks are used. Figures 4a, 4b and 4c show how τ_t dominates in all cases. The assignations t^l and c^l are computed for each frame with $\Delta\tau = 4$ as given by (10) and (11), and the sums C and T are then obtained using (12) and (13). Figure 5 shows the ratio $r_{classification}$ for the masks ζ_{soft} (blue) and ζ_{hard} (red) as the energy level E_0 increases. The ratio $r_{detection}$ is shown for the masks ζ_{soft} (green) and ζ_{hard} (black). As the energy level E_0 increases, the ratio $r_{classification}$ in (14) increases to reach 1, and the ratio (15) decreases to reach 0.

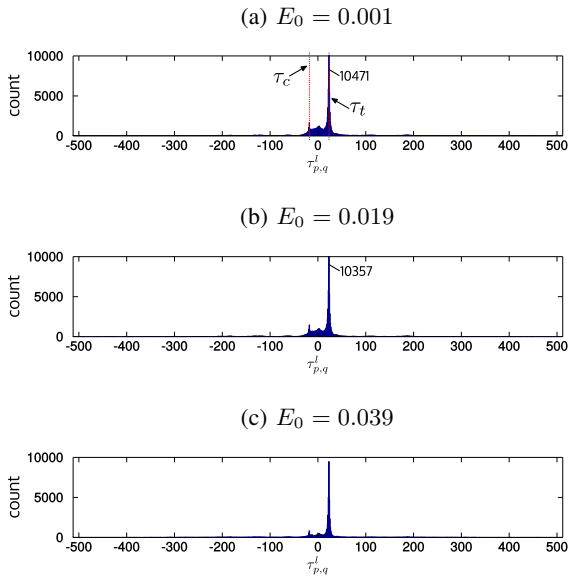


Fig. 4: $\tau^l_{p,q}$ distribution according to E_0 with ζ_{hard}.

To evaluate performance, we define u according to (21) by being equal to $r_{detection}$ when $r_{classification}$ reaches 95%, which is close to the ideal 100%, while still providing a significant value for $r_{detection}$. The larger the value of u, the more robust the method is to broadband coherent noise. For the computational load, (22) evaluates the ratio of frames that are processed ($z^l = 1$) over the total number of frames L. A frame is processed with mask ζ_{hard} only when the condition in (23) is satisfied, while all frames are processed with mask ζ_{soft}. When all frames are processed, the computational load is at 100%.

$$u = r_{detection}(E_0) \text{ when } r_{classification}(E_0) = 0.95 \quad (21)$$

$$\text{computational load} = \frac{1}{L}\sum_{l=0}^{L-1} z^l \quad (22)$$

$$z^l = \begin{cases} 1 & \theta^l_p \geq \theta_{min}, \theta^l_q \geq \theta_{min} \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

Figure 5 shows the results obtained. The values of u with the masks ζ_{soft} and ζ_{hard} are 2.8%, and 13.1%, respectively.

Performance for a wide range of noise and reverberation levels is also investigated. Table III presents values of u when the masks ζ_{soft} and ζ_{hard} are used. The parameter θ_{min} takes different values to assess its impact on performance and computation load (shown in parenthesis). The value of u always improves for ζ_{hard} with $\theta_{min} = 0$ when compared to ζ_{soft} (in the best case, the gain in u reaches 16.1% for RT60 = 0, SNR = 0 dB). It is also interesting to note that ζ_{hard} with $\theta_{min} = 10$, when compared to ζ_{soft}, provides superior performance for u and reduces the computational load down to 23% in the best case.

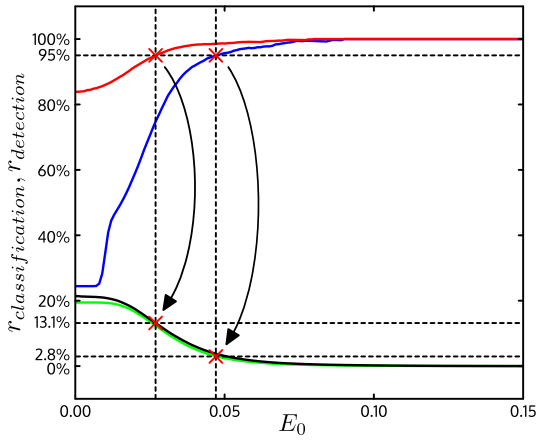


Fig. 5: TDOA estimation for a simulation with RT60 = 500 msec and SNR = 10 dB: $r_{classification}$ as a function of E_0 for ζ_{soft} (blue) and ζ_{hard} (red), and $r_{detection}$ as a function of E_0 for ζ_{soft} (green) and ζ_{hard} (black)

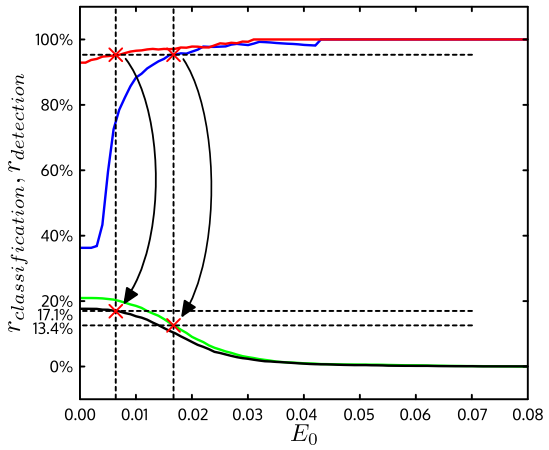


Fig. 6: TDOA estimation on a mobile robot: $r_{classification}$ as a function of E_0 for ζ_{soft} (blue) and ζ_{hard} (red), and $r_{detection}$ as a function of E_0 for ζ_{soft} (green) and ζ_{hard} (black)

2) Mobile Robot: A robot equiped with an 8-microphone array is used to validate performance of the approach in a real-life environment. For this experiment, only the two microphones fixed on the front of the robot torso are used to measure performance. A male speaker speaks at 2 meters on one side of the robot and a loudspeaker streams a stationary white noise on the other side. The reverberation level in the room is measured to be RT60 = 800 msec, and the SNR is 8.1 dB for the microphone on the male speaker side, and 7.3 dB for the second microphone close to the loudspeaker. The $r_{detection}$ and $r_{classification}$ ratios are shown in Fig. 6 for masks ζ_{soft} and ζ_{hard}. u is equal to 13.4% and 17.1% for the ζ_{soft} and ζ_{hard}, respectively. The variable θ_{min} is set to 5 and the computational load is estimated to be at 60%. This suggests that the proposed mask is more robust to coherent broadband noise, and reduces computational load.

TABLE III: Values of u and computational load (in parenthesis) with masks ζ_{soft} and ζ_{hard} in relation to θ_{min}

RT60 (msec)	SNR (dB)	ζ_{soft}	ζ_{hard}				
			$\theta_{min}=0$	$\theta_{min}=10$	$\theta_{min}=20$	$\theta_{min}=30$	$\theta_{min}=40$
0	20	47.3% (100%)	**54.5%** (100%)	**53.3%** (56%)	**48.3%** (49%)	40.3% (40%)	32.0% (32%)
	15	39.1% (100%)	**49.6%** (100%)	**47.8%** (51%)	**39.3%** (40%)	28.4% (28%)	19.2% (19%)
	10	29.5% (100%)	**43.0%** (100%)	**40.2%** (44%)	28.1% (28%)	16.2% (16%)	8.8% (9%)
	5	19.6% (100%)	**35.5%** (100%)	**30.8%** (35%)	16.6% (17%)	7.0% (7%)	2.8% (3%)
	0	10.7% (100%)	**26.8%** (100%)	**20.0%** (23%)	7.2% (7%)	1.8% (2%)	0.4% (0%)
250	20	36.2% (100%)	**46.3%** (100%)	**46.0%** (62%)	**44.3%** (54%)	**39.3%** (45%)	32.6% (36%)
	15	26.8% (100%)	**40.6%** (100%)	**40.1%** (56%)	**36.0%** (44%)	**27.9%** (31%)	19.6% (21%)
	10	17.5% (100%)	**33.5%** (100%)	**32.5%** (48%)	**25.5%** (30%)	15.7% (17%)	9.0% (10%)
	5	9.4% (100%)	**25.8%** (100%)	**23.9%** (36%)	**14.5%** (17%)	6.5% (7%)	2.7% (3%)
	0	4.3% (100%)	**17.6%** (100%)	**14.5%** (23%)	**6.0%** (7%)	1.6% (2%)	0.4% (0%)
500	20	13.6% (100%)	**24.3%** (100%)	**24.3%** (66%)	**24.3%** (62%)	**24.0%** (54%)	**22.2%** (45%)
	15	7.3% (100%)	**18.6%** (100%)	**18.6%** (63%)	**18.6%** (53%)	**17.2%** (40%)	**14.1%** (28%)
	10	2.8% (100%)	**13.1%** (100%)	**13.1%** (56%)	**13.0%** (38%)	**9.9%** (22%)	**6.7%** (12%)
	5	0.8% (100%)	**9.0%** (100%)	**9.0%** (43%)	**8.2%** (21%)	**4.4%** (9%)	**1.9%** (3%)
	0	0.2% (100%)	**5.2%** (100%)	**5.2%** (26%)	**3.5%** (8%)	**1.0%** (2%)	0.2% (0%)
750	20	4.5% (100%)	**9.3%** (100%)	**9.3%** (68%)	**9.3%** (64%)	**9.3%** (59%)	**9.2%** (50%)
	15	1.6% (100%)	**5.9%** (100%)	**5.9%** (65%)	**5.9%** (57%)	**5.9%** (44%)	**5.8%** (32%)
	10	0.4% (100%)	**3.4%** (100%)	**3.4%** (59%)	**3.4%** (42%)	**2.9%** (25%)	**2.5%** (14%)
	5	0.0% (100%)	**2.1%** (100%)	**2.1%** (47%)	**2.1%** (23%)	**1.5%** (10%)	**0.7%** (4%)
	0	0.0% (100%)	**1.2%** (100%)	**1.2%** (29%)	**1.1%** (9%)	**0.4%** (2%)	0.0% (0%)
1000	20	2.1% (100%)	**4.3%** (100%)	**4.3%** (69%)	**4.3%** (66%)	**4.3%** (60%)	**4.3%** (52%)
	15	0.7% (100%)	**2.1%** (100%)	**2.1%** (66%)	**2.1%** (59%)	**2.1%** (47%)	**2.1%** (34%)
	10	0.2% (100%)	**1.0%** (100%)	**1.0%** (61%)	**1.0%** (45%)	**1.0%** (28%)	**1.0%** (16%)
	5	0.0% (100%)	**0.2%** (100%)	**0.2%** (49%)	**0.2%** (25%)	**0.2%** (10%)	**0.1%** (4%)
	0	0.0% (100%)	**0.3%** (100%)	**0.3%** (30%)	**0.3%** (9%)	**0.2%** (2%)	0.0% (0%)

IV. Conclusion

This paper presents a TDOA estimation method based on the weighted GCC-PHAT that is robust to coherent broadband noise. The approach is based on a new hard mask that improves noise robustness and reduces the computational load of the system. In future work, we plan to integrate the use of this TDOA estimation method with a sound source localization system such as the one proposed in ManyEars. We also plan to have the parameters automatically adapt to the environment to find optimal localization performance in diversified conditions.

References

[1] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2005, pp. 4040–4045.

[2] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 742 – 752, 2007.

[3] F. Grondin and F. Michaud, "WISS, a speaker identification system for mobile robots," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2012, pp. 1817–1822.

[4] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2009, pp. 2033–2038.

[5] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, vol. 2, 2003, pp. 1228–1233.

[6] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2006.

[7] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Autonomous Robots*, pp. 217–232, 2013.

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

[9] C. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2009, pp. 2027–2032.

[10] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization and its application to multimodal human tracking," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2011, pp. 143–148.

[11] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2012, pp. 694–699.

[12] P. Danes and J. Bonnal, "Information-theoretic detection of broadband sources in a coherent beamspace music scheme," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2010, pp. 1976–1981.

[13] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2012, pp. 2625–2628.

[14] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Proc. of the IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2012, pp. 521–524.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[16] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[17] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.