Noise Mask for TDOA Sound Source Localization of Speech on Mobile Robots in Noisy Environments

François Grondin and François Michaud

***Abstract*— Sound source localization is an important challenge for mobile robots operating in real life settings. Sound sources of interest, such as speech, are often corrupted by broadband coherent noise sound source(s) that are non-stationary during transitions between steady-state segments. The interfering noise introduces localization ambiguities leading to the localization of invalid sound sources. Masks to reduce such interferences perform well under stationary noise, but the performance degrades as localization of invalid sound sources generated by noise appear and disappear suddenly during transitions between steady-state. This paper presents a new mask based on speech non-stationarity to discriminate between the time difference of arrival (TDOA) of speech source and noise transition. Simulations and experiments on a mobile robot suggest that the proposed technique improve TDOA discrimination and reduces significantly localization of invalid sound sources caused by noise.**

I. Introduction

Sound source localization on mobile robots can be performed using Multiple Signal Classification based on Standard Eigenvalue Decomposition (SEVD-MUSIC). This method was initially used for narrowband signals [1], and was adapted for broadband sound source localization [2]. This method improves noise robustness by decomposing the covariance matrix obtained from the spectral observations at each microphone in both noise and noisy signal subspaces. The direction of a sound source corresponds to the directional vector orthogonal to the noise subspace. This method performs well as long as the noise power is weaker than the sound source of interest. Multiple Signal Classification based on Generalized Eigenvalue Decomposition (GEVD-MUSIC) is proposed to deal with this issue [3]. SEVD-MUSIC and GEVD-MUSIC significantly improve robustness to noise, but remains sensitive to reverberation, and eigenvalue decomposition involves a high computational load. Multiple Signal Classification based on Generalized Singular Value Decomposition (GSVD-MUSIC) is proposed to reduce computational load [4]. While this method remains sensitive to highly reverberant environments, it improves localization accuracy because eigenvectors are mutually orthogonal.

Another approach for sound source localization on mobile robots is to use a delay-and-sum beamformer [5]. This method relies on the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) to perform time difference of arrival (TDOA) estimation. This method is robust to reverberation but remains sensitive to broadband additive noise. Beamforming based on weighted GCC-PHAT, as implemented in the ManyEars framework [6], [7], is appealing to perform sound source localization: it is robust to reverberation, has low-complexity, and is robust to additive noise if noise can be masked. Noise masking highly relies on an accurate background noise estimation. To improve robustness to noise of GCC-PHAT, a non-recursive unbounded mask is derived from the instantaneous signal-to-noise ratio (SNR), and the results is correlated with the intensity of the sound source of interest [8]. To make the result independent of the amplitude of the sound source, a recursive soft mask with bounded values is used in ManyEars [6], [7]. We also recently introduced in ManyEars a hard binary mask to improve its robustness to broadband coherent noise [9].

The masks implemented so far in ManyEars [8], [9] rely on the Minima Controlled Recursive Averaging (MCRA) method to estimate background noise [10], [11]. However, background noise usually changes over time if the robot operates in dynamic environments. Computer fans on the robot can also start and stop according to the central processing unit (CPU) usage. When noise changes, the MCRA method requires a few seconds to adapt and to provide an accurate noise estimation. During this transition, the masks no longer hide noise components, which can lead to localization of noise sources instead of speech. To our knowledge, this problematic has not yet been addressed in other works.

To overcome this problem, this paper presents a new mask that can deal with noise transition. This mask, which we refer to as transition mask, is an extension of the hard mask previously introduced [9] and measures the stationarity of previous and future frames to mask noisy time-frequency regions. The paper is organized as follows. Section II briefly presents the weighted GCC-PHAT method and noise masks used for background noise estimation, to then explain our new transition noise mask. Section III describes the experiments conducted in simulation and on a mobile robot platform.

II. Weighted GCC-PHAT with Noise Masks

To implement GCC-PHAT, the Short-Time Fourier Transform (STFT) $X^l_m[k]$ is computed using (1) for each microphone m with frames of N samples, and spaced by a hop of ΔN samples. The expression $x_m[n]$ represents the signal from each microphone and n is the discrete-time index. A Hann window $w[n]$ of N samples is used to reduce spectral

This work was not 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, Interdisciplinary Institute for Technological Innovation, 3000 boul. de l'Université, Québec (Canada) J1K 0A5, {Francois.Grondin2,Francois.Michaud}@USherbrooke.ca

leakage. The variables l and k stand for the frame and bin indexes, respectively.

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

The weighted GCC-PHAT between microphones p and q is expressed by (2). The expression $(...)^*$ stands for the complex conjugate operator, while the variable $\zeta_{pq}^l[k]$ is a frequency mask introduced to reduce coherent noise contribution. The expression ϵ is also added to avoid overflow when the spectrum magnitude goes to zero. The weighted GCC-PHAT result is $R_{pq}^l[n]$.

$$R_{pq}^l[n] = \frac{1}{N} \sum_{k=0}^{N-1} \frac{\zeta_{pq}^l[k] X_p^l[k] X_q^l[k]^*}{|X_p^l[k]||X_q^l[k]| + \epsilon} e^{j2\pi kn/N} \quad (2)$$

To estimate TDOA, the cross-correlation result R_{pq}^l lies in the interval $[0, N-1]$ and is mapped on the new interval $[-N/2+1, N/2]$ using the expression $\hat{R}_{pq}^l$ as in (3). The modulo N operation ensures that a negative TDOA value, initially mapped to a value greater than $N/2$ with the Inverse Fourier Transform (IFFT) in (2), is brought back to the desired range.

$$\hat{R}_{pq}^l[n] = R_{pq}^l\left[n \bmod N\right] \quad (3)$$

The maximum cross-correlation index is obtained by (4) to provide the plausible range of a TDOA value, which lies in the interval $[-n_{pq}^{max}, n_{pq}^{max}]$.

$$\tau_{pq}^l = \arg\max_n \left(\hat{R}_{pq}^l[n]\right) \quad -n_{pq}^{max} \leq n \leq n_{pq}^{max} \quad (4)$$

The maximum cross-correlation index corresponds to the case when the sound source and the pair of microphones lie on the same plane, and is given by (5). The variables f_s, c, $\mathbf{x}_p$ and $\mathbf{x}_q$ stand for the sampling rate (in samples/sec), the speed of sound (in m/sec), and the positions of microphones p and q, respectively.

$$n_{pq}^{max} = \left(\frac{f_s}{c}\right) \|\mathbf{x}_p - \mathbf{x}_q\| \quad (5)$$

The maximum cross-correlation energy E_{pq}^l is given by (6).

$$E_{pq}^l = \hat{R}_{pq}^l[\tau_{pq}^l] \quad (6)$$

To reduce noise contribution, a soft mask introduced in [8] relies on the *a priori* SNR, obtained from the background noise estimated with the MCRA method. A mask is estimated for each microphone channel and both masks are combined to obtain $\zeta_{pq}^l[k]$ defined by (7).

$$\zeta_{pq}^l[k] = \zeta_p^l[k] \zeta_q^l[k] \quad (7)$$

When this mask is used, broadband coherent noise may leak from each bin and generate undesired dominant peak in the TDOA estimation result. To solve this issue, a binary mask for each microphone channel, called hard mask because it only takes discrete values of 0 and 1, can be used with the MCRA method [9].

A. Transition Noise Mask

The MCRA method is efficient to estimate stationary noise as long as it can take a few seconds to adapt to the background noise. When a robot operates in a dynamic environment, noise is often non-stationary during sound transitions. For instance, fans installed on the robots or in the room may start and stop as time goes by, and this introduces abrupt transitions between steady-state segments.

To consider such conditions, we define two working hypotheses [12]:

1) Speech phonemes are stationary for less than 25 msec.
2) After a fast transition, noise stays stationary for more than 25 msec.

The transition noise mask exploits these characteristics to differentiate noise from speech. The mask is computed using the product of $|X_p^l[k]|^2$ and $|X_q^l[k]|^2$, as opposed to the soft and hard masks that are computed directly from the individual spectrum of each microphone. The expression $|X_{pq}^l[k]|^2 = |X_p^l[k]|^2 |X_q^l[k]|^2$ improves speech-noise discrimination as the inter-channel spectrum magnitude correlation is higher for speech than for noise. A rectangular window smooths the log magnitude of the spectrum frequency-wise according to (8). The smoothed result is expressed by $Y_{pq}^l[k]$, and the rectangular window is made of $(2W+1)$ samples. The spectral envelope is extracted with the logarithm and this operation also reduces the spectral distortion caused by the high energy outliers when the moving average is computed. The expression ϵ is added to avoid underflow when the spectrum power goes to zero. This operation reduces the variance of the log magnitude between adjacent frames.

$$Y_{pq}^l[k] = \frac{1}{2W+1} \sum_{\Delta k=-W}^{W} \log\left(|X_{pq}^l[k+\Delta k]|^2 + \epsilon\right) \quad (8)$$

$Y_{pq}^l[k]$ is used to evaluate two parameters:

- $A_{pq}^l[k]$, the difference between the current power level and the minimum value of a buffer made of the ΔA previous frames, using (9):

$$A_{pq}^l[k] = Y_{pq}^l[k] - \min\left\{Y_{pq}^{l-\Delta A+1}[k], \ldots, Y_{pq}^l[k]\right\} \quad (9)$$

- $B_{pq}^l[k]$, the difference between the current power value and the minimum value within a window of ΔB future frames, defined by (10):

$$B_{pq}^l[k] = Y_{pq}^l[k] - \min\left\{Y_{pq}^l[k], \ldots, Y_{pq}^{l+\Delta B-1}[k]\right\} \quad (10)$$

This introduces a latency of $\Delta N \Delta B / f_s$ sec because the window requires samples ahead in time.

To explain the role of these two parameters, Fig. 1 illustrates their influences with four types of signal: 1) steady state noise (NSS); 2) noise level rises quickly (NRI); 3) noise level drops quickly (NDP); 4) speech (SPH). Fig. 1a) to 1d) illustrate how $Y^l_{pq}[k]$ reduces the power variance. For $A^l_{pq}[k]$, Fig. 1f) and 1h) illustrates that large differences are observed for NRI and SPH, while small differences are observed for the NSS and NPD shown in Fig. 1e) and 1g). For $B^l_{pq}[k]$, Fig. 1i) and 1j) shows that small differences are observed for NSS and NRI, and Fig. 1k) and 1l) illustrates that differences are greater for the NDP and SPH.

These observations indicate that only speech generates large values for both the signals $A^l_{pq}[k]$ and $B^l_{pq}[k]$. This condition, defined as $D^l_{pq}[k]$, can be identified using (11), with θ_A and θ_B proposed as fixed threshold values.

$$D^l_{pq}[k] = \begin{cases} 1 & \left(A^l_{pq}[k] > \theta_A\right) \wedge \left(B^l_{pq}[k] > \theta_B\right) \\ 0 & otherwise \end{cases} \tag{11}$$

The transition noise mask $\zeta^l_{pq}[k]$ can therefore be defined in relation to $D^l_{pq}[k]$ as in (12) and used in (2) during GCC-PHAT computation. A majority vote is performed in (12), with ΔD being the window size in frames and θ_D the majority threshold, to filter out false detections. For instance, $D^l_{pq}[k]$ may sporadically trigger a false value of 1 when non-stationary percussive sounds are observed.

$$\zeta^l_{pq}[k] = \begin{cases} 1 & \left(\displaystyle\sum_{l'=l}^{l+\Delta D} D^{l'}_{pq}[k]\right) > \theta_D \\ 0 & otherwise \end{cases} \tag{12}$$

III. Results

In this section, performance using the soft, hard and transition noise masks are presented for experiments conducted in simulation and with a mobile robot. Table I presents the parameters used. Only W, ΔA, ΔB, ΔD, θ_A, θ_B, and θ_D have to be fine tuned empirically, the others are being set according to environmental settings. f_s is set at 48 kHz for the acquisition of high frequency components in speech. N is the same as in [9] and the hop size ΔN ensures an overlap of 50%. The speed of sound c is defined at 20°C and 101.1 kPa [13]. The expression ϵ is chosen to avoid overflow or underflow while still be small enough to preserve the precision of (2) and (8). The expression W is set to a value that provides a smoothing window that preserves frequency resolution. ΔA and ΔB are chosen to exploit the non-stationarity of speech, and also deal with reverberation. ΔD and θ_A, θ_B and θ_D are chosen empirically to capture most of the speech features and reject noisy regions. With the proposed parameters, a latency of $\Delta N \Delta B / f_s = 213$ msec is introduced. A latency of this range can be considered negligible in human-robot vocal interaction.

The theoretical delays τ_t and τ_c are associated to the target (i.e., the sound source of interest) and the noise sources, respectively. Positions of the target source and the noise source in relation to the microphones are known for the experiments conducted, allowing to evaluate the performance of TDOA sound localization using the different noise masks. τ_t and τ_c are defined by (13) and (14), where $\mathbf{s}_t$ and $\mathbf{s}_c$ are vectors which hold the cartesian positions of the sound source of interest and noise sources, respectively.

TABLE I: Parameters used in the experiments

Parameters	Values	Parameters	Values
N	1024	ΔA	20
ΔN	512	ΔB	20
f_s (Hz)	48000	ΔD	3
c	343.0	θ_A	2.0
ϵ	1E-10	θ_B	2.0
W	10	θ_D	1

$$\tau_t = \left(\frac{f_s}{c}\right)\left(\|\mathbf{s}_t - \mathbf{x}_p\| - \|\mathbf{s}_t - \mathbf{x}_q\|\right) \tag{13}$$

$$\tau_c = \left(\frac{f_s}{c}\right)\left(\|\mathbf{s}_c - \mathbf{x}_p\| - \|\mathbf{s}_c - \mathbf{x}_q\|\right) \tag{14}$$

To characterize the performance of the transition noise mask, the following metrics are introduced. The expression t^l_{pq} takes a value of 1 when the delay τ^l_{pq} is assigned to the sound source of interest, and a value of 0 otherwise. A delay τ^l_{pq} is considered to be assigned to the sound source of interest when the absolute value of the difference with the theoretical delay τ_t is less or equal to a constant $\Delta\tau$, as shown in (15).

$$t^l_{pq} = \begin{cases} 0 & \left|\tau^l_{pq} - \tau_t\right| > \Delta\tau \\ 1 & \left|\tau^l_{pq} - \tau_t\right| \leq \Delta\tau \end{cases} \tag{15}$$

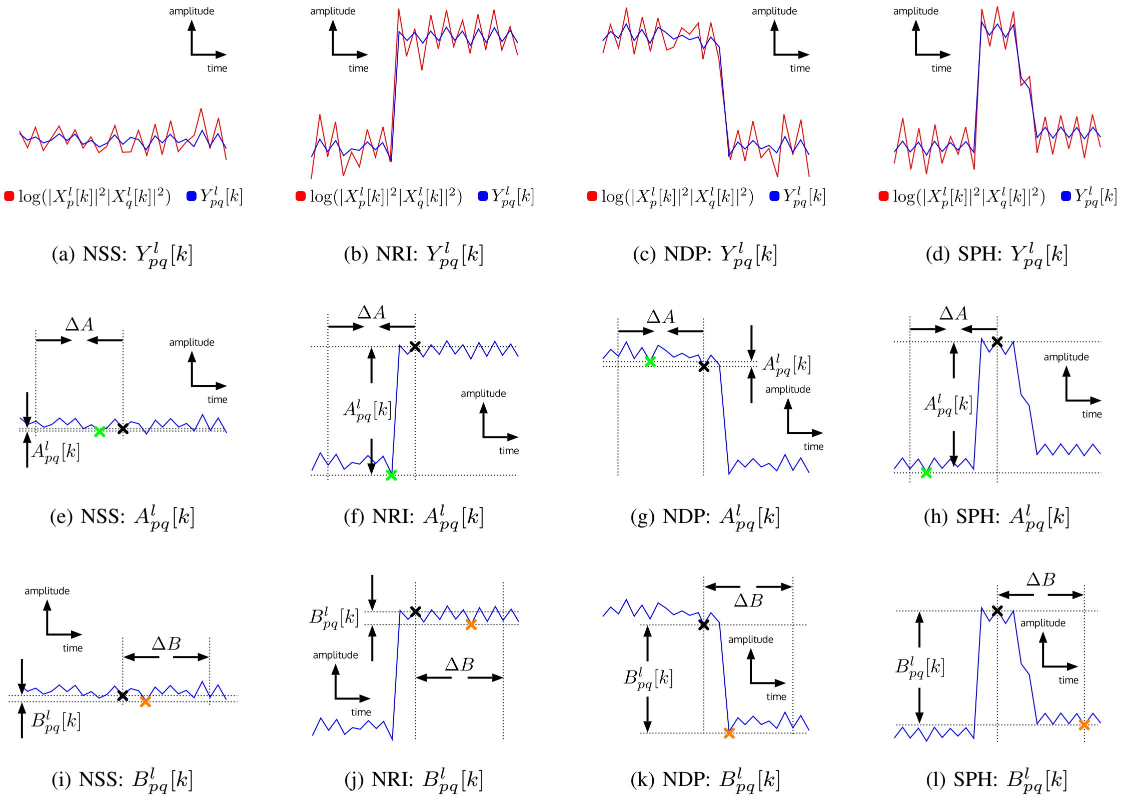
Similarly, the expression c^l_{pq} is set to 1 when the noise source generates the coherence noise delay, and a value of 0 otherwise. When coherent noise dominates, the TDOA is assigned to the noise source and the absolute value of the difference with the delay τ_c is less or equal to $\Delta\tau$, provided by (16).

$$c^l_{pq} = \begin{cases} 0 & \left|\tau^l_{pq} - \tau_c\right| > \Delta\tau \\ 1 & \left|\tau^l_{pq} - \tau_c\right| \leq \Delta\tau \end{cases} \tag{16}$$

A higher energy level normally indicates a higher confidence level, which is usually used with a subsequent confidence-based decision stage [7]. To evaluate energy levels, t^l_{pq} and c^l_{pq} are multiplied by the associated maximum cross-correlation energy E^l_{pq}, as given by (17) and (18). These weighted averages give more importance to delays associated to a high energy level, and thus provide meaningful metrics to measure overall performance.

$$T_{pq} = \left(\sum_{l=0}^{L-1} t^l_{pq} E^l_{pq}\right) \Big/ \left(\sum_{l=0}^{L-1} E^l_{pq}\right) \tag{17}$$

$$C_{pq} = \left(\sum_{l=0}^{L-1} c^l_{pq} E^l_{pq}\right) \Big/ \left(\sum_{l=0}^{L-1} E^l_{pq}\right) \tag{18}$$



(a) NSS: $Y_{pq}^l[k]$ (b) NRI: $Y_{pq}^l[k]$ (c) NDP: $Y_{pq}^l[k]$ (d) SPH: $Y_{pq}^l[k]$

(e) NSS: $A_{pq}^l[k]$ (f) NRI: $A_{pq}^l[k]$ (g) NDP: $A_{pq}^l[k]$ (h) SPH: $A_{pq}^l[k]$

(i) NSS: $B_{pq}^l[k]$ (j) NRI: $B_{pq}^l[k]$ (k) NDP: $B_{pq}^l[k]$ (l) SPH: $B_{pq}^l[k]$

Fig. 1: Signals $Y_m^l[k]$, $A_m^l[k]$ and $B_m^l[k]$ for NSS, NRI, NDP and SPH

TABLE II: Simulation parameters

Parameters	Values
$\mathbf{x}_1$	$-0.1\hat{\imath} - 0.1\hat{\jmath} + 0.0\hat{k}$
$\mathbf{x}_2$	$-0.1\hat{\imath} + 0.1\hat{\jmath} + 0.0\hat{k}$
$\mathbf{s}_t$	$+2.0\hat{\imath} + 4.0\hat{\jmath} + 0.5\hat{k}$
$\mathbf{s}_c$	$-2.0\hat{\imath} + 0.0\hat{\jmath} + 0.1\hat{k}$
$\Delta\tau$	5

A. *Simulation Results*

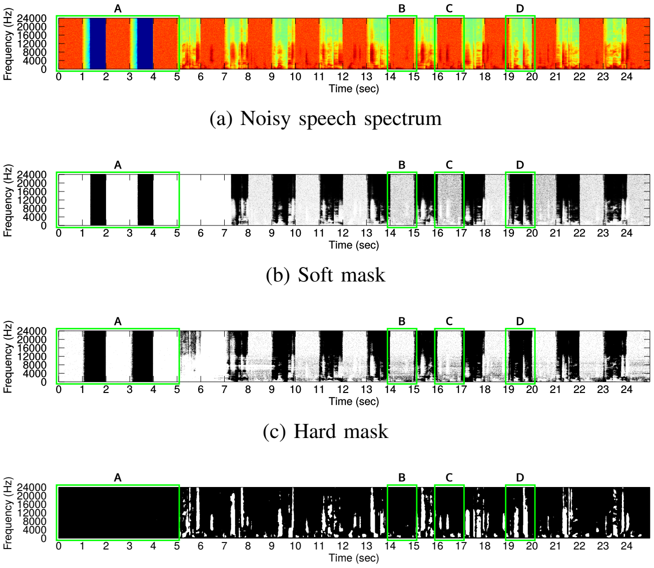
Simulations are performed under a wide range of reverberation levels and SNRs. Speech is used as the sound source of interest, and the noise source is made of white noise bursts. These sources are convolved with their respective room impulse responses (RIRs) generated for a room of dimensions 10 m × 6 m × 3 m using Allen and Berkley image method [14].

Table II presents the parameters used for the simulations. The unit vectors $\hat{\imath}$, $\hat{\jmath}$ and $\hat{k}$ represent the x, y and z axes of a three-dimensional cartesian coordinate system. The vector $\hat{k}$ points toward the ceiling of the simulated room. The expression $\Delta\tau$ is chosen to be large enough to capture delays close to the theoretical value. The simulations are performed with more than 43 minutes of speech, i.e., 243600 frames.

Figure 2 illustrates the noisy speech spectrum and the soft, hard and transition noise masks, when the reverberation time (RT60) is 250 msec and the SNR is 10 dB. Noise bursts are active for one second, and spaced by periods of one second of silence. The soft and hard masks erroneously capture noise bursts in segments A, B and C. The transition mask rejects noise bursts in segments A and B, and leave the noisy speech unaltered in segment C for the regions where speech is more powerful than noise, as desired. For all masks, the clean speech segment D is left unaltered, as desired.

Figure 3 illustrates the distribution of τ_t and τ_c for each type of masks, when the energy level E_{pq}^l is greater than zero. The soft and hard masks lead to a distribution peak at the noise delay τ_c, while the distribution obtained with the transition mask reaches its maximum value at the target delay τ_t.

Table III presents T_{pq} and C_{pq} for a wide range of reverberation levels and SNRs. The sum of T_{pq} and C_{pq} may not equal to 100% if some delay values are not assigned to the sound source of interest nor to the coherent noise source. In the ideal case, T_{pq} should reach 100%, and C_{pq} 0%. Results indicate that the transition mask provide better performance for all reverberation levels and SNR values. With low reverberation (RT60 = 0 msec) and high signal-to-noise ratio (SNR = 20 dB), T_{pq} and C_{pq} reach 99% and 0%, respectively, using the transition mask, compared $T_{pq} = 69\%$ and $C_{pq} = 31\%$ with the hard mask and $T_{pq} = 43\%$ and $C_{pq} = 57\%$ using the soft mask. In high reverberation conditions (RT60 = 500 msec) and low signal-to-noise ratio (SNR = 0 dB), the transition mask outperforms the hard and soft masks, with $T_{pq} = 34\%$ and $C_{pq} = 10\%$ compared to $T_{pq} = 1\%$ and $C_{pq} = 98\%$ using the hard mask and



(a) Noisy speech spectrum

(b) Soft mask

(c) Hard mask

(d) Transition mask

Fig. 2: Sound spectrum with RT60 = 250 msec and SNR = 10 dB. In a), high and low power levels are shown in red and blue respectively. In b) to d), the black and white colors stand for the values 0 and 1, respectively.

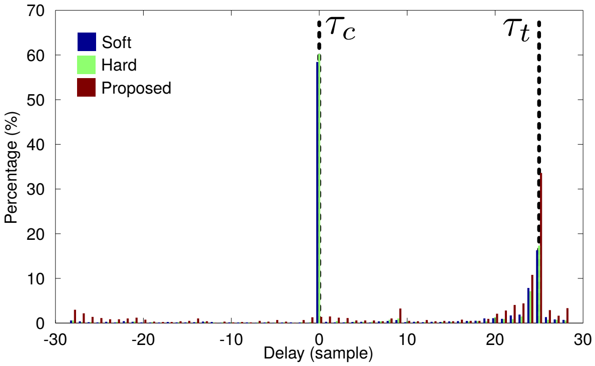


Fig. 3: Delay distribution generated with the soft, hard and transition masks for simulation with RT60 = 250 msec and SNR = 10 dB

$T_{pq} = 2\%$ and $C_{pq} = 97\%$ using the soft mask.

B. Experiments using a Mobile Robot

Figure 4 shows the IRL-1 robot equipped with an 8-microphone array. IRL-1 was used to validate the performance of the approach in a real-life environment. To scale down the complexity of the experiment and to simplify analysis, only two microphones located on the front of the IRL-1 are used. Since ManyEars relies on the sum of the weighted GCC-PHAT between each pair of microphones, the proposed method with two microphones can easily be adapted to an 8-microphone approach such as ManyEars.

TABLE III: T_{pq} and C_{pq} (in parenthesis) using the soft, hard and transition masks

RT60 (msec)	SNR (dB)	Soft Mask	Hard Mask	**Transition Mask**
0	20	43% (57%)	69% (31%)	**99% (0%)**
	15	39% (61%)	45% (55%)	**99% (1%)**
	10	30% (70%)	22% (78%)	**99% (1%)**
	5	14% (86%)	11% (89%)	**98% (1%)**
	0	11% (89%)	10% (90%)	**99% (1%)**
200	20	28% (70%)	55% (44%)	**93% (1%)**
	15	25% (74%)	32% (66%)	**93% (1%)**
	10	18% (82%)	12% (88%)	**92% (2%)**
	5	8% (92%)	6% (94%)	**91% (2%)**
	0	6% (94%)	5% (95%)	**92% (2%)**
300	20	17% (79%)	40% (51%)	**69% (4%)**
	15	15% (81%)	21% (74%)	**69% (4%)**
	10	10% (87%)	6% (93%)	**68% (5%)**
	5	4% (95%)	3% (96%)	**66% (6%)**
	0	3% (96%)	2% (97%)	**66% (6%)**
400	20	13% (80%)	33% (51%)	**48% (7%)**
	15	11% (82%)	16% (76%)	**48% (7%)**
	10	8% (88%)	4% (93%)	**47% (7%)**
	5	3% (95%)	2% (97%)	**45% (8%)**
	0	2% (97%)	2% (98%)	**46% (8%)**
500	20	10% (81%)	28% (50%)	**34% (9%)**
	15	9% (83%)	13% (76%)	**35% (9%)**
	10	7% (88%)	3% (94%)	**35% (9%)**
	5	3% (95%)	2% (97%)	**34% (9%)**
	0	2% (97%)	1% (98%)	**34% (10%)**

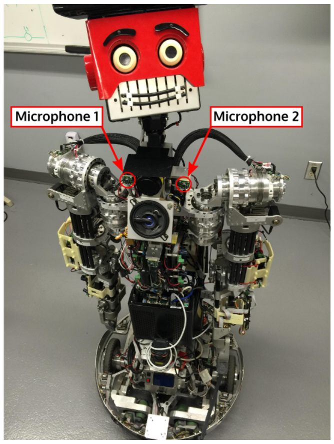
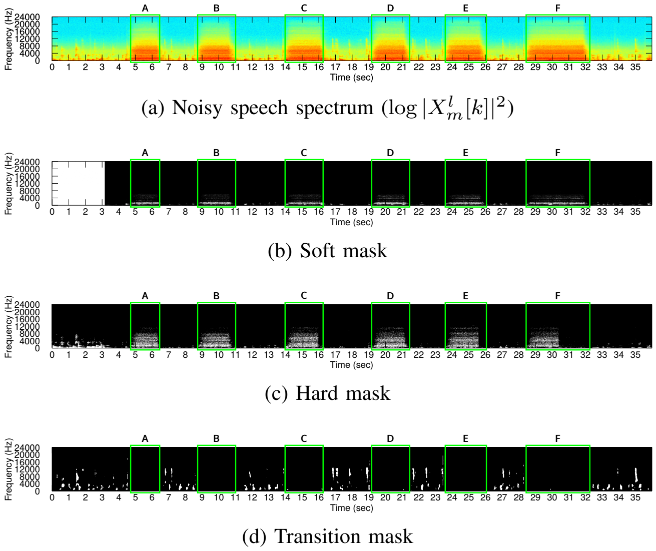


Fig. 4: Pair of microphones used on the IRL-1 robot

Male speech is played by a loudspeaker installed on the left of the robot, and a noisy hairdryer is turned on and off by a participant standing on the right of the robot. The reverberation level in the room is RT60 = 800 msec. When the hairdryer is active, the SNR is -3.3 dB for the microphone on the loudspeaker side, and -6.3 dB for the second microphone on the hairdryer side.

Figure 5a presents the signal spectrum for this experiment. It shows that the hairdryer dominates speech in segments A, B, C, D, E and F, and that the soft and hard masks erroneously capture the hairdryer noise. The noise is dominant over speech, and therefore the transition mask reject all time-frequency regions when the noise is active, and captures speech when noise is inactive.

Figure 6 illustrates the distribution of τ_t and τ_c obtained



(a) Noisy speech spectrum ($\log |X^l_m[k]|^2$)

(b) Soft mask

(c) Hard mask

(d) Transition mask

Fig. 5: Signal spectrum for the experiment conducted using IRL-1. High and low power levels are shown in red and blue respectively, in 5a. The black and white colors stand for the values 0 and 1, respectively, in 5b, 5c and 5d.

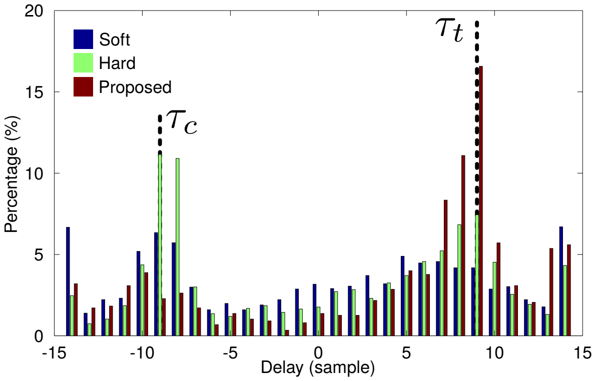


Fig. 6: Delay distribution generated using the soft, hard and transition masks on the IRL-1 robot

with each type of masks, when the energy level E^l_{pq} is not zero. Using the soft or the hard masks lead to a distribution peak at the noise delay τ_c, while the distribution obtained with the transition mask reaches its maximum value at the target delay τ_t.

The soft mask generates weighted averages of $T_{pq} = 55\%$ and $C_{pq} = 34\%$, and the hard mask produces weighted averages of $T_{pq} = 38\%$ and $C_{pq} = 53\%$. Using the transition mask results in $T_{pq} = 79\%$ and $C_{pq} = 16\%$, significantly improving robustness to noise compared to the soft and hard masks.

IV. Conclusion

This paper presents a transition noise mask for TDOA estimation based on the weighted GCC-PHAT to improve robustness to coherent broadband noise with abrupt transitions. The transition noise mask relies on the non-stationarity of speech, and results indicate that it outperforms soft and hard masks used by the weighted GCC-PHAT approach. It requires the introduction of a small latency, which reveals to be an acceptable trade-off to avoid invalid localization of sound sources generated by noise: for instance, the presence of invalid sound sources could lead to inappropriate responses of the robot (such as reorienting its head in direction of the loudest sound source) if additional processing is required to identify invalid sound sources as noise.

The next step with this work is to integrate the use of this transition noise mask with the implementation of the weighted GCC-PHAT in ManyEars. We also plan to use the information provided by the transition mask to improve computation in ManyEars, for instance by pausing when most frequency bins in the transition mask equal to 0.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[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] 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*, 2003, pp. 1228–1233.

[9] F. Grondin and F. Michaud, “Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2015.

[10] 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.

[11] 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.

[12] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer Science & Business Media, 2007.

[13] A. Zuckerwar, *Handbook of the Speed of Sound in Real Gases*. Elsevier Science, 2002.

[14] 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.