IMPROVEMENT OF ACOUSTIC LOCALIZATION USING A SHORT TIME SPECTRAL ATTENUATION WITH A NOVEL SUPPRESSION RULE

Daniele Salvati
AVIRES Lab,
University of Udine
Udine, Italy
kabit@tiscali.it

Sergio Canazza
AVIRES Lab,
University of Udine
Udine, Italy
sergio.canazza@uniud.it

ABSTRACT

This paper proposes innovative de-noise filters in a framework, whose aim is the localization of an acoustic source in a noisy environment. The main focuses are the automatic detection of transient sound events and the separation of the events of interest from the noise. A microphone array is used to capture time-spatial information and an adaptive filter can be initialized to learn the ambient noise spectrum when signals of interest are absent. We propose an algorithm based on the Short Time Spectral Attenuation method to remove the noise from each sensor of the array, before the source localization task is performed. The Time Difference Of Arrival (TDOA) methods are used for multiple sources localization. The experimental results show the efficiency of our framework in stationary noisy environments.

1. INTRODUCTION

Microphone array signal processing is used to extract useful spatial information from acoustic signals. The most important applications of such beam-forming, spatial sound-field recording and direction of arrival (DOA) estimation (based on TDOA method) are potentially used to solve the problem of single or multiple source localizations, estimation of number of sources, sources separation, de-reverberation and echo reduction [1]. Many techniques have been developed for microphone array processing in different acoustic environments: reverberant or free-field, far-field or near-field. Our interest concerns events detection and DOA estimation in a free-field environment. The question of DOA involves two main steps. The first is TDOA estimation, based on the measurement of the time difference between the signals received by different microphones. The second locates the source position processing the TDOA information with the knowledge of sensor geometry array and acoustic environment. In an array design, the multiple microphones setup depends on the application. The most common configurations are linear, circular and rectangular, and correct spacing between each sensor is important to avoid spatial aliasing problems, in relation with the spectra of the acoustic sources. The signal model received by the *n*th microphone in a free-field environment with multiple sources is:

$$y_n(k) = \sum_{m=1}^{M} \alpha_{nm} \cdot s_m[k - t_m - \tau_{nm}] + v(k) \qquad (1)$$

where M is the total number of the signals, α_{nm} is the attenuation of the sound propagation (inversely proportional to the distance from source m to microphone n), s_m are the unknown source signals, t_m is the propagation time from the *n*th microphone to the sensor reference, τ_{nm} is the TDOA of the *m*th signal between the *n*th microphone and the reference, *v(k)* is ambient noise. The problem of the acoustic source automatic detection is discriminating should consider *v(k)*. A possible approach is based on the development of a system capable of recognizing the sound source [2] in the case of screams and gunshots. This technique builds a set of audio features, which derive from sound analysis, in order to detect unique characteristics of some particular audio event. The results of this approach are very interesting, at least in the specific acoustic environment considered.

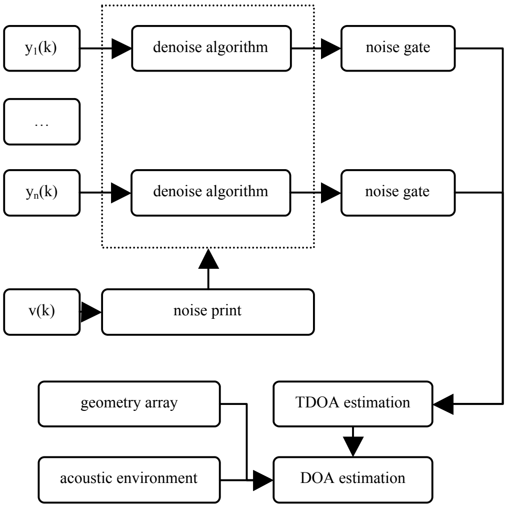

Figure 1: *The architecture of the processor, showing the data flow of all the tasks of the experimental system implementation.*

Our approach is based on trying to remove *v(k)* from (1), instead. The de-noising algorithm employed is an evolution of the Ephraim and Malah suppression rule (EMSR [3] and [4]), and it is denominated CMSR (Canazza-Mian Suppression Rule). It is a frequency method filter that requires to perform a noise estimation. To perform the best denoising process, the noise signal should be stationary during the acoustic event. However, since this condition hardly occurs in real life, after the denoising processing that cleans the residual noise as good as possible, a noise gate should be applied to each microphone. The noise gate removes the parts of the signal below the selected threshold, which is set to a very low level. The architecture of the processor is summarized in Figure 1 and shows the data flow of all the tasks of the experimental system implementation.

In the following section we present the TDOA estimation. In Sec. 3 the (classic) noise reduction algorithms are introduced, with the Ephraim-Malah Suppression Rule (EMSR). In Sec. 4 new suppression rules are proposed. Finally in Sec. 5 the qualities of the proposed method are discussed, showing some experimental results.

2. ACOUSTIC SOURCE LOCALIZATION

The generalized cross-correlation (GCC) [5] is the most common technique employed for TDOA estimation; it is used with spatially separated microphone pairs. The relative time delay τ_{12} is obtained by an estimation of the peaks detector in the filter cross-correlation function:

$$\hat{\tau}_{12} = \arg\max_{\tau} r_{y_1 y_2}(\tau) \tag{2}$$

where τ is the time lag and $r_{y_1 y_2}(\tau)$ is the GCC function:

$$r_{y_1 y_2}(\tau) = \sum_{k=0}^{N-1} \Psi(f) \cdot S_{y_1 y_2}(f) \cdot e^{\frac{j2\pi\tau f}{N}} \tag{3}$$

where N is the number of samples of the observation time, $\psi(f)$ is the frequency domain weighting function, and the cross-spectrum of the two signals is defined as:

$$S_{y_1 y_2}(k) = E\left\{Y_1(f) \cdot Y_2^*(f)\right\} \tag{4}$$

where $Y_i(f)$ (i = 1, 2) is the Discrete Fourier Transform (DFT) of the signal and * denotes complex conjugate. GCC is used for minimizing the influence of uncorrelated noise and interference, and maximizing the peak in correspondence of the time delay. In the free-field environment with moderate noise this method works very well and is computationally efficient, but in low signal to noise ratio (SNR) levels a clear distinction between signals of interest and noise is not achieved. This is another question to be pointed out in order to highlight the usefulness of a de-noising filter, not only for the separation of the acoustic events from the noise, but also for a better resolution of the localization with a low SNR. In our experimental results we show the GCC with different weighting functions proposed by Knapp and Carter. The simple cross-correlation (CC) is computed when $\psi_{cc} = 1$. The CC is estimated using the DFT and the inverse DFT (IDFT), which can be efficiently implemented with the fast Fourier transformer (FFT).

The *Roth* weighting function is calculated according to the SNR value of one signal. The GCC estimates the impulse response of the optimum linear (Wiener-Hopf) filter:

$$\Psi_{ROTH}(f) = \frac{1}{S_{y_1 y_1}(f)} \tag{5}$$

The *Roth* processor has the desirable effect of suppressing the frequency regions in which the power spectral noise density is large and the estimate of the cross power spectral signal density is likely to be in error.

The *Smoothed Coherence Transform* (SCOT) assigns weighting according to the SNR of both signals:

$$\Psi_{SCOT}(f) = \frac{1}{\sqrt{S_{y_1 y_1}(f) \cdot S_{y_2 y_2}(f)}} \tag{6}$$

The *Phase Transform* (PHAT) function normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC:

$$\Psi_{PHAT}(f) = \frac{1}{\left|S_{y_1 y_2}(f)\right|} \tag{7}$$

The *Hannan & Thomson* (HT) weighting function, also known as Maximum-Likelihood (ML) GCC, is:

$$\Psi_{HT}(f) = \frac{\left|\gamma_{12}(f)\right|^2}{\left|S_{y_1 y_2}(f)\right| \cdot (1 - \left|\gamma_{12}(f)\right|^2)} \tag{8}$$

where $|\gamma_{12}|^2$ is the magnitude square coherence (MSC) function between y_1 and y_2:

$$\left|\gamma_{12}(f)\right|^2 = \frac{\left|S_{12}(f)\right|^2}{S_{y_1 y_1}(f) \cdot S_{y_2 y_2}(f)} \tag{9}$$

When several microphone pairs are available, the source position can be estimated as the point in space that best fits a set of TDOA measurements. Comparison methods are analyzed in [6].

Consequently the source localization is a DOA estimation. In our system, we chose the most simple array configuration (uniform linear array, ULA) to analyze pre-filter denoising effects, and we consider the parametric model of a far-field environment. The DOA value θ is calculated as:

$$\theta = \arcsin\left(\frac{\tau \cdot c}{d}\right) \tag{10}$$

where c is the speed of sound, and d the distance between microphones. The assumed DOA range is: -90° +90°, where zero is in front of the array.

3. DE-NOISE FILTER

Over the last ten years research in the audio restoration field has focused on the planning of algorithms, which subtend a plurality of models and hypotheses on the sound reality and have been developed in connection with the peculiar problem which the system intends to solve:

- Restoration can aim at retrieving the intelligibility of the spoken parts during a communication occurring in a perturbed environment, where the critical factor is real-time working, even if at the expense of a high loss of vocal timbre quality (communication between pilots and control tower, between divers and the mother ship, or between troops on enemy soil);
- A commercial objective can, instead, concern the understanding of the spoken parts during a communication in a noisy environment. In this case, real-time working of the system is essential, but with an at least partial preservation of the original timbre: communication between *mobile* devices in shopping centres, or at

parties, concerts and in traffic jam, where the useful signal (speech) to noise is very low.

- Intelligibility retrieval is also the aim of restoration in the forensic field. In this case, real-time working is not required, but faithfulness to timbre must be guaranteed for the identification of the vocal print.
- Restoration of musical recordings, which must offer satisfying solutions to the problems connected with the time-varying feature peculiar to musical signals. Real time is not required (the work of the restorer often requires 10 or 20 times the real time). In this case the noise reduction interventions could: 1) concern only the cases in which the internal evidence of the degradation is unquestionable, without going beyond the technological level of that time; 2) aims at a commercial edition; 3) have the purpose of obtaining a historical reconstruction of the recording as it were listened to at the time.
- This essay considers scenarios in which the real time is a critical factor, with a preservation of the original timbre (useful for audio source recognition, to be performed in a possible next step) and where a hypothesis on the useful signal (voice or noise) cannot be made.

The specific methodologies of audio restoration can be schematized in at least three different categories, according to the information used by the algorithm during the phase of noise attenuation.

1) *Frequency methods;* these algorithms require that the operator have little information to carry out the restoration (a *priori* information): only an estimate of the noise present is necessary (noise print), since it is assumed to be stationary along the entire signal. Any further information needed (a *posteriori* information) is automatically calculated by the restoration software through the analysis of the characteristics of the signal. Since these algorithms are easy to use and are generally applied to different typologies of audio signals, they are employed in commercial hardware and software systems.

2) Algorithms in the time domain, which use *signal models*; *a priori* information is employed to estimate the probable distribution of the sound events, the excitation signal and the filter coefficients. Therefore the algorithm carries out (a *posteriori* information) the signal tracking. The models, which can be applied to different signal typologies, are "non-informative" (they have little *a priori* information): it is therefore necessary to detail the model from time to time, according to the signal being examined.

3) Restoration through *analysis to synthesis* and restoration based on *source models*; in this case only *a priori* information is required. It is to be found in the knowledge we have about the system that produced the audio document and the analysis of the sound material.

Our scenario (real time, no any *a priori* information available) suggests the use of the Frequency methods, which are based on the Short Time Spectral Attenuation. These de-noise systems consist of two important components: a noise estimation method and a suppression rule. In this section we focus on the suppression rule. We do not discuss the estimation method, which is equally important for the final noise reduction system; however, our method can readily be combined with any existing estimation method. In Sec. 3.1. we present an overview of the STSA methods. Sec. 3.2. introduces the EMSR method ([3] and [4]).

• 3.1. Frequency domain methods

These techniques employ a signal analysis through the Short-Time Fourier Transform (which is calculated on small partially overlapped portions of the signal: STFT) and can be considered as a non-stationary adaptation of the Wiener filter [7] in the frequency domain. In particular, Short Time Spectral Attenuation (STSA) consists in applying the short-time spectrum of the noise to a time-varying suppression and does not require the definition of a model for the audio signal.

Suppose considering the useful signal *x(t)* as a stationary aleatory process to which some noise *z(t)* is added (uncorrelated with *x(t)*) to produce the degraded signal *y(t)*:

$$y(t) = x(t) + z(t) \tag{11}$$

The relation that connects the respective power spectral densities is therefore:

$$P_y(\omega) = P_x(\omega) + P_z(\omega) \tag{12}$$

with ω the frequency index.

If we hypothesize to succeed in retrieving an adequate estimate of $P_z(\omega)$, during the silence intervals of the signal *y(t),* and in the musical portions that of $P_y(\omega)$, we can expect to obtain an estimate of the spectrum of *x(t)* by subtracting $P_z(\omega)$ from $P_y(\omega)$; the initial assumption of stationariness can be considered locally satisfied since short temporal windows are employed.

Note that the use of a short-time signal analysis is equivalent to the use of a filter bank. First each channel (that is, the output of each filter) is appropriately attenuated and then it is possible to proceed with the synthesis of the restored signal. The time-varying attenuation applied to each channel is calculated through a determined suppression rule, which has the purpose to produce an estimate (for each channel) of the noise power. Each particular STSA technique is characterised by the implementation of the filter bank and of the suppression rule.

Often the short-time analysis is carried out through the STFT ([3], [8] and [9]). In [10], instead, non-linear filter banks are introduced.

Historically, the STSA methodology was developed during the '70s in order to remove the noise in the transmission of spoken parts. The new STSA techniques for audio restoration are an adaptation of these first elaborations. Traditionally, the interpretation as STFT is a notion deriving from the analysis of the spoken parts. The phase remains an open problem: in the STFT interpretation, the attenuation corresponds to a change of the short-time spectrum magnitude only. The opinion that the phase does not need to be processed owing to the psycho-acoustic properties of the human ear is widespread. Indeed, the "insensitivity to the phase" of the human ear is proved only in the case of stationary audio signals and for the Fourier Transform phase. On the contrary, in the case of the STFT phase, variations among subsequent short-time frames can cause audible effects (such as frequency modulation). It is important to highlight that in the classic STSA techniques the possibility to process the phase does not exist, since no hypothesis is made on the characteristics of the audio signal.

If we denote the STFT of the *y(t)* noisy signal with *Y*(*t, ω_k*), where *t* represents the temporal index and ω_k the frequency index (with *K* = 1,…, *N*: *N* represents the number of STFT channels), the result of the suppressing rule application can be interpreted as the application of a *G*(*t, ω_k*) gain to each value *Y*(*t, ω_k*) of the STFT of the noisy signal . This gain corresponds to a signal attenuation and is included between 0 and 1.

In most of the suppression rules, G(*t, ω_k*) only depends on the noisy signal power level (measured in the same point) $|\mathrm{Y}(\mathrm{t},\ \omega_\mathrm{k})|^2$ and on the estimate of the noisy power at the ω_k frequency:

$$\hat{P}_z(\omega_k) = E\left\{|Z(t,\omega_k)|^2\right\} \tag{13}$$

(which does not depend on the temporal index *t* due to the presumed noise stationariness). At this point a *relative* signal can be defined:

$$Q(t,\omega_k) = \frac{|Y(t,\omega_k)|^2}{\hat{P}_z(\omega_k)} \tag{14}$$

which, starting from the hypothesis that the *z*(*t*) noise is not correlated to the *x(t)* signal, we deduce should be greater than 1:

$$E\{Q(t,\omega_k)\} = 1 + \frac{E\left\{|X(t,\omega_k)|^2\right\}}{\hat{P}_z(\omega_k)} \tag{15}$$

A typical suppression rule is based on the Wiener filter [7] and can be formulated as follows:

$$G(t,\omega_k) = \frac{|Y(t,\omega_k)|^2 - \hat{P}_z(\omega_k)}{|Y(t,\omega_k)|^2} \tag{16}$$

Another rule, called Power-Subtraction, is illustrated in [11]. Comparing the characteristics of the two rules in connection with the relative signal $Q(t,\omega_k)$, we deduce that these suppression rules share the same behaviour:

- $G(t,\omega_k) = 1$, where the relative signal is high ($Q(t,\omega_k) >> 1$)
- $\lim\limits_{Q(t,\omega_k)\to 1} G(t,\omega_k) = 0$.

  That is, the gain tends to 0 in the case in which only the noise is present (relative signal equal to 1). In this sense, in some cases an overvaluation of the estimated noise power is used.

Other more elaborated suppression rules depend on both the relative signal and on a priori knowledge of the corrupted signal, that is on a priori knowledge of the probability distribution of the in-band signals [11] or on the signal to noise ratio [3]. Usually, the mistake made by these procedures in retrieving the original sound spectrum has an audible effect, since the difference between the spectral densities can give a negative result at some frequencies. Should we decide to arbitrarily force the negative results to zero, in the final signal there will be a disturbance, constituted of numerous random frequency pseudo-sinusoids, which start and finish in a rapid succession, generating what in literature is known as *musical noise* [3].

• 3.2. EMSR method

After the Wiener solution, many variants, which are also affected by *musical noise*, even if in a minor way, were proposed. On the contrary a substantial progress was made with the solution hereinafter proposed. The work, carried out in [3] and [4], aims at minimising the mean square error (MSE) in the estimation of the spectral components (Fourier coefficients) of the musical signal, of which A_k indicates the magnitude:

$$E\left\{\left(A_k - \hat{A}_k\right)^2\right\} \tag{17}$$

By modelling A_k as a statistically independent null mean Gaussian aleatory variables, the obtained solution is:

$$\hat{A}_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right]Y_k \tag{18}$$

where:

$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$; $\gamma_k = \frac{Y_k^{\ 2}}{E\left[|Z_k|^2\right]}$ (*a posteriori* SNR); $\xi_k = \frac{E\left[|X_k|^2\right]}{E\left[|Z_k|^2\right]}$ (*a priori* SNR).

and X_k, Z_k, Y_k are the spectral components of the clean signal *x(t),* of the noise *z(t)* and of the noisy signal *y(t)* respectively.

I_0 and I_1 are the Bessel modified functions of the zero and first order. Note that the quantity ξ_k can only be estimated, since the clean signal is not available. The estimate calculation is developed according to two models, one based on a maximum likelihood approach and the other based on a decision directed approach. Since the latter one turned out to be the best, we report it here (*n* is the frame index):

$$\hat{\xi}_k(n) = \alpha\frac{\hat{X}_k(n-1)}{Z_k(n-1)} + (1-\alpha)P\left[y_k(n) - 1\right],\ 0 \le \alpha < 1 \tag{19}$$

where: $P[x] = \begin{cases} x & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$

In [9] the behaviour of the filter based on such an estimator is analysed; after a notation change the gain applied to each spectral component *k* to the *p*-th frame is:

$$G(k,p) = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+Y_{post}(k,p)}\right)\left(\frac{Y_{prio}(k,p)}{1+Y_{prio}(k,p)}\right)}\cdot$$
$$\cdot M\left[\left(1+Y_{post}(k,p)\right)\left(\frac{Y_{prio}(k,p)}{1+Y_{prio}(k,p)}\right)\right] \tag{20}$$

where: $M[\vartheta] = \exp\left(-\frac{\vartheta}{2}\right)\left[(1+\vartheta)I_0\left(\frac{\vartheta}{2}\right) + \vartheta I_1\left(\frac{\vartheta}{2}\right)\right]$

where the two parameters Y_{post} and Y_{prio} are calculated as:

$$Y_{post}(k,p) = \frac{|X(k,p)|^2}{v(k)} - 1$$

$$Y_{prio}(k,p) = (1-\alpha)P\Big[Y_{post}(k,p)\Big] + \alpha\frac{|G(k,p-1)X(k,p-1)|^2}{v(k)}$$

where: $P[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

where *v(k)* is the noise power at the *k* frequency. The *α* parameter controls the balance between the current frame information and that of the preceding one. By varying this parameter, the filter smoothing effect can be regulated. Y_{prio} has less variance than Y_{post}: in this way, it is less probable that a musical noise occurs. However, it is important to point out that when α increases, the response delay is even higher than the transients; therefore there is a low-pass effect on the occasion of rapid signal attacks.

A good adaptation to the non-stationary noise case [9] is another advantage of the proposed algorithm; this is particularly important in our scenario, where the noise can be generated by wind, water, rain, walking, etc., and all these events can change in a medium time scale.

We also have to point out that by increasing the overlapping of the analysis windows, the statistic correlation degree between the frames increases. This results in a limitation of the noise reduction power of the filter. A detailed analysis carried out by the authors shows that a hard overlapping (exceeding 80%) can give acceptable results only by increasing the value of *α*.

In [4] an evolution of the EMSR suppression rule, based on an estimator which minimizes the logarithmic mean square error in the estimation of the signal spectral components, was presented, and that is:

$$E\left\{\left(\log A_k - \log\hat{A}_k\right)^2\right\} \qquad (21)$$

where A_k represents the module of the k-th Fourier coefficient. The estimator obtained in [4] is:

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}Y_k \qquad (22)$$

This realization represents a significant evolution of the standard EMSR: in particular it produces less musical noise at the expense of a minor uniformity (however it is hardly audible due to the minor remaining noise).

4. INNOVATIVE FILTERS BASED ON EMSR

We present the implementation of three filters created by the authors, which represent an evolution of the Ephraim and Malah suppression rule.

The first one, CMSR (Canazza-Mian Suppression Rule[1]), is based on the idea of using a "punctual" suppression without memory, (Wiener-like), in the case of a null estimate of Y_{post}; the pseudo-code is the following:

```
IF  $Y_{post}(k,p) > 0$
                    $\alpha = 0.98$
ELSE
                    $\alpha = 0$
END
```

The experiments carried out confirm that the filter performs very well, with a greater noise rejection than the classic EMSR and it has the prerogative of not introducing musical noise. Furthermore the behaviour in the transients follows that of EMSR without having the impression of a "low-pass filter" application.

The second algorithm, called CMSRα, takes into account the information of the last two frames. More precisely, before brutally setting the parameter at zero, we observe if the preceding frame also contained a null Y_{post}. The pseudo-code is:

```
IF  $Y_{post}(k,p) > 0$
  $\alpha = 0.98$ ;
  $Y_{prio}(k,p) = (1-\alpha)P[Y_{post}(k,p)] + \alpha\frac{|G(k,p-1)Y(k,p-1)|^2}{v(k)}$ ;
ELSE
  IF  $Y_{post}(k,p-1) > 0$
    $\alpha = 0.98$ ;
    $Y_{prio}(k,p) = (1-\alpha)P[Y_{post}(k,p-1)] + \alpha\frac{|G(k,p-2)Y(k,p-2)|^2}{v(k)}$ ;
    $Y_{post}(k,p) = Y_{post}(k,p-1)$ ;
  ELSE
    $\alpha = 0$
  END
END
```

The last algorithm, called CMSRβ, calculates Y_{prio} as follows:

$$\begin{aligned} &Y_{prio}(k,p) = \\ &= 0.98\cdot\left[(1-\alpha)P[Y_{post}(k,p)] + \alpha\frac{|G(k,p-1)Y(k,p-1)|^2}{v(k)}\right] + \\ &+0.02\cdot\left[(1-\alpha)P[Y_{post}(k,p-1)] + \alpha\frac{|G(k,p-2)Y(k,p-2)|^2}{v(k)}\right] \end{aligned} \qquad (23)$$

In order to estimate Y_{prio}, the decisional strategy of using the past previous frames p-1 at 98% and p-2 at 2% (only as corrective parameter) was adopted.

[1] Gian Antonio Mian (1942-2006), co-author of this paper, was a professor of Digital Signal Processing at the Dept. of Information Engineering, University of Padova, a leading researcher and an outstanding teacher whose brightness and kindness we will always remember. This Suppression Rule is affectionately dedicated to his memory.

5. EXPERIMENTAL RESULTS

We performed a detailed validation of the methods described above. In particular we tested our algorithms (see Sec. 4) in comparison with the Wiener Filter, the Power-Subtraction and the EMSR method. We carried out the test in a real (very noisy) scenario with a time-variant SNR (5÷60 dB). Each filter is able to periodically update the noise print in an automatic way.

From the experimental results it is evident that in the signals processed by Wiener Filter there is a significant quantity of residual noise modulated around the signal frequencies. Furthermore, a strong presence of musical noise can be observed. The EMSR filter has decidedly less musical noise, but it is necessary to regulate the parameter α in order to find a good compromise between noise attenuation and the transient distortion. The CMSR filter enables/allows a higher noise attenuation without causing distortions. CMSRα and CMSRβ show a minor noise reduction compared to the CMRS. In addition, the analysis also showed a delay in reaction in the sound events decay (transition from “only signal” to “only noise”).

Important indications can be drawn from the tests we carried out: two fundamental parameters of the dimension and of the analysis of the windows overlapping were changed. A high overlapping, exceeding 80%, brings no advantages to the noise reduction. On the contrary, the advantage brought to the window dimension increase is evident, provided that the overlapping increases as well (to make the next step constant in terms of samples number). However, a larger window has the disadvantage of leaving a grater quantity of residual noise. The tests show that dimensions of 4096 to 8192 samples represent a good compromise (a sampling frequency of 44.1 kHz was considered).

Figure 2 shows the gain trend introduced by each filter at the varying of the noisy signal SNR in a real scenario. The term gain indicates the difference between the de-noised signal SNR and the input signal SNR. For all input SNR’s, the CMSR has a good performance.

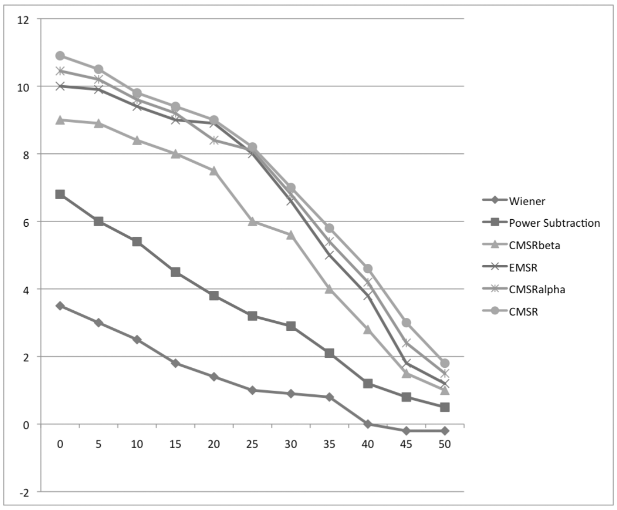

Figure 2: *Gain trend introduced by the filters in the frequency domain at the varying of the input SNR (y-axis: SNR$_{out}$-SNR$_{in}$; x-axis: SNR$_{in}$ – both in dB). The best gain of the CMSR filter can be observed for all the SNR$_{in}$.*

The next figures show the experimental results of the implementation system (see Fig. 1). The improvement of acoustic source localization using de-noise filter, which is the CMSR in accordance with the results in Fig. 2, is highlighted to compare Fig. 3, 4, 5, 6 (events in environmental noise) with Fig. 7, which displays the process of the TDOA estimation with the de-noise and the noise gate filter. The maximum SNR measured in the analyzed transient acoustic event is 20 dB. The peak of the capture snapshot in Fig. 7, which appears at the arrival of sound, shows the signal of interest alone during the whole acoustic event, proving a complete removal of the noise.

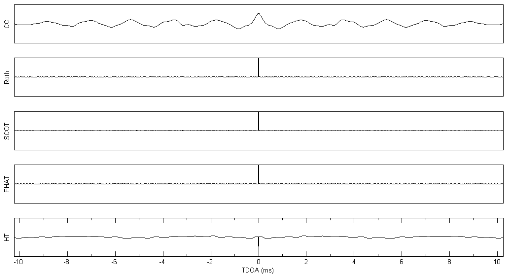

Figure 3: *Environmental noise – GCC function in presence of the noise only. In this case the automatic system detection can’t discriminate it from a new event (x axis: time lag in ms). In this situation the step of catching the noise print can be initialized to learn the spectral noise.*

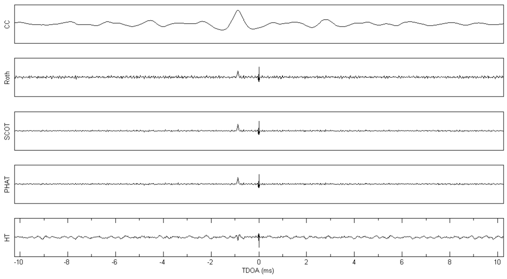

Figure 4: *Environmental noise – an acoustic signal is arriving to the microphone array and two peaks are observed. SNR < 5dB.*

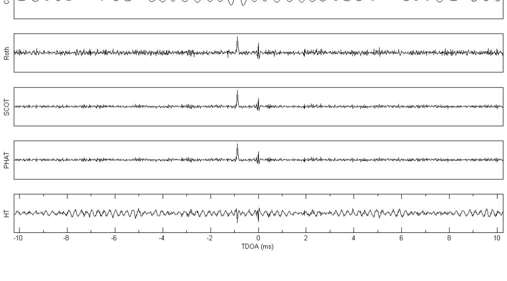

Figure 5: *Environmental noise – SNR < 5 dB; two peaks are observed.*

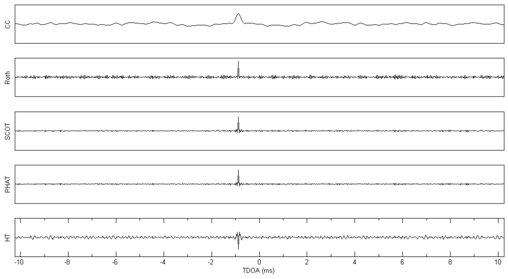

Figure 6: *Environmental noise – SNR > 5. The energy of the signal increases and the GCC function hides the noise peak.*

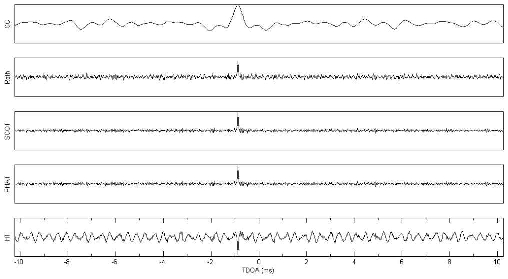

Figure 7: *De-noise – Until a signal of interest is detected, the TDOA processor does not display any peak, as it uses a de-noise filter (CMSR) and a noise gate (in this case with a threshold set to -65 dB). The system can discriminate the arrival of any new event. The figure shows a snapshot of the same acoustic event presented in the previous illustrations with de-noise task. During the entire performance (from 0 to 20 dB of SNR) we have one peak, until it disappears at the end of the acoustic wave.*

6. CONCLUSIONS

We presented a framework for audio source localization, where a noise reduction algorithm is integrated. The main purpose was to enhance the well-known EMSR filter. We studied and optimized the performance of the most used de-noise algorithms in terms of signal-to-noise ratio, which is a simple but limited quality measure. We showed a comparison among the methods and we included the best filter, called CMSR, in our framework. In a real noisy scenario, our test presented the TDOA estimation using GCC method with different weighting function, and the ability of the CMSR filter to separate a sound event from a stationary noise, to improve the detection and localization of acoustic sources.

ACKNOWLEDGMENTS

This work was partially supported by the Italian Ministry of University and Scientific Research within the framework of the project “Dynamic and unstructured environments interpretation, virtualization and monitoring by an integrated autonomous system of sensors and robots”, 2009-2012.

REFERENCES

[1] J. Benesty, J. Chen and Y. Huang, *Microphone Array Signal Processing*, Springer, Volume 1, 2008.

[2] G. Valenzise, L. Gerosa and M. Tagliasacchi, F. Antonacci, A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems”, in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance,* Sept. 2007, pp 21-26.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[4] Y. Ephraim and D. Malah, “Speech Enhancement using a minimum mean-square error log-spectral amplitude estimator,” in *IEEE Transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443-445, 1985.

[5] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.

[6] A Brutti, M. Omologo and P. Svaizer, “Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection” in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp.69-72.

[7] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*, Cambridge, MIT Press, Massachusetts, 1949.

[8] J.S. Lim and A.V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” in *Proceedings IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[9] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” in *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345-349, 1994.

[10] R.J. McAulay and M.L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137-145, 1980.

[11] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in IEEE Transactions on *Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.