Statistical underwater noise level estimation for marine mammals whistle detection

Sylvain Busson, Cedric Gervaise

10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Statistical underwater noise level estimation for marine mammal whistle detection

Sylvain Busson, Cédric Gervaise
ENSIETA, DTN, 2 rue F. Verny, 29206 Brest {sylvain.busson,cedric.gervaise}@ensieta.fr

Passive Acoustic Monitoring (PAM) of marine mammal vocalizations has been intensively used in applications such as marine wildlife surveys. Marine mammals regularly produce sounds to echolocate, to communicate and while foraging and PAM is a complementary tool of visual-based observations for density estimation. Vocalizations of marine mammals are divided into two categories: impulse sounds and frequency modulation whistles. A common way to detect whistles consists in forming signal spectrogram and to check for frequency modulation tracks. Whistles tracks are made of energetic pixels identified by comparing each spectrogram pixel levels with an estimate of local noise. Despite simple to implement and quite effective, usual noise level estimation via signal low-pass filtering lead to false alarms in non-stationary noise conditions as in the vicinity of human activities or in coastal observatories. A noise level estimation algorithm is proposed by taking into account the statistical properties of noise only samples of the spectrogram. The algorithm is based on minima statistics of a time-frequency neighborhood of the current spectrogram pixel where noise only pixels follow a chi-square distribution. The estimation is effectively applied on field data. Major drawbacks coming from low-pass filtering are avoided and the noise level is in closer agreement to the real noise level. Interpolation schemes are proposed to keep the real-time implementation of the algorithm feasible.

1 Introduction

Studies on impacts and mitigations of anthropogenic sounds on marine mammals are of growing importance [1] and Passive Acoustic Monitoring (PAM) [2] is a tool of choice for such studies. Marine mammals regularly produce sounds to echolocate, to communicate and while foraging. PAM is a complementary tool of visual-based observations. Vocalizations of marine mammals can be divided into two categories: impulse sounds and frequency modulation whistles. The first stage of PAM system is dedicated to the detection of bioacoustics activities. A common way to detect underwater bioacoustics signals consists in computing signal spectrogram. A binary test is applied on each spectrogram pixel to discriminate between H_0 = 'noise only' and H_1 = 'noise + signal' hypothesis. The decision is based on the energy of the pixel: if the energy exceeds a threshold the decision 'signal + noise' is made:

$$\begin{cases} H_0 : \gamma_S < T \\ H_1 : \gamma_S > T \end{cases} \quad (1)$$

where γ_S is the pixel energy and T is the energy threshold. Alternatively, equation (1) can be expressed in terms of signal to noise ratio (SNR) by defining $T = \lambda\gamma_n$ where γ_n is an estimation of the 'noise only' energy level. If H_0 is decided when a signal is present, it is called a miss and when H_1 = is decided when a signal is absent, it is a false alarm.

One common implementation of the noise level estimation (NLE) is computed via time averaging obtained either by low-pass filtering or smoothing techniques. A local NLE is computed using few pixels in the same frequency channel around the current pixel. Despite simple to implement and quite effective, this approach could lead to major drawbacks. The integration time should be chosen carefully by taking into account the signal stationarity. In one hand, if the signal is composed of strong energetic impulse sounds, like echolocation clicks of marine mammals, the noise level estimation will rise after the impulse and this will result in a period of overestimation. During that signal shadowing period, the PAM system will not be able to detect any weaker signals. In a group of marine mammals, clicks and whistles can occur simultaneously. As clicks carry far more energy than whistles, they will not be detected a during shadowing period. In the other hand, if the time constant is chosen too short in non-stationary noise conditions as it occurs in the vicinity of human activities or in coastal observatories, the noise level will be underestimated. In that case, the detection will be prone to a high false alarm rate.

We propose a NLE algorithm by taking into account not only few pixels before the current pixel but hundreds of pixels of a time-frequency neighborhood centered on the current pixel. Using the average energy of pixel in the neighborhood could lead to overestimate the noise level if signal is present. In our algorithm, the estimation is based on the statistics of noise only pixels which are represented by the lowest energy pixels of the neighborhood. The parameters of the algorithm are the shape and the size of the neighborhood and the number of the pixels of weak energy taken into account for the estimation of noise statistics. The noise level is obtained

by an approximation of statistics of the distribution the pixels of the spectrogram which are assumed to follow a chi-square distribution.
The estimation is effectively applied on both synthetic and field data. Major drawbacks coming from low-pass filtering are avoided. The statistical NLE has weaker biais and variance. However, this noise level estimation technique is time-consuming. Interpolation techniques are proposed to keep the real-time implementation of the algorithm feasible. A 10 seconds signal with a sampling frequency of 44100 Hz, which corresponds to a 9.10^5 pixels spectrogram, is processed in no more than 1.5 seconds on a standard computer.

2 Methods

2.1 Probability distribution of spectrogram coefficients

Let be $x[n]$ a discrete signal and $X_\omega[n,k]$ its Short-Time Fourier Transform (STFT) composed of successive discrete Fourier Transforms of portions of $x[n]$. Each portions are weighted by an analysis window $\omega[n]$ of length M. The spectrogram of $x[n]$, noted $S_x^\omega[n,k]$, corresponds to the sum of the square modulus of $X_\omega^r[n,k]$ and $X_\omega^i[n,k]$, the STFT real and imaginary parts respectively:

$$S_x^\omega[n,k] = X_\omega^r[n,k]^2 + X_\omega^i[n,k]^2, \tag{2}$$

$$X_\omega^r[n,k] = \sum_{m=0}^{M-1} x[n-m]\omega[m]\cos(-2\pi k\frac{m}{M}), \tag{3}$$

$$X_\omega^i[n,k] = \sum_{m=0}^{M-1} x[n-m]\omega[m]\sin(-2\pi k\frac{m}{M}), \tag{4}$$

Under the assumption that $x[n]$ is a white noise represented by a stationary Gaussian random process with variance γ_n^2, $X_\omega^r[n,k]$ and $X_\omega^i[n,k]$ follow a Gaussian law and $S_x^\omega[n,k]$ is proportional to a χ^2 variable with the degrees of freedom δ, a non-centrality parameter θ and a coefficient of proportionality equal to α [3]:

$$p_{\chi^2}(x) = \frac{1}{(2\alpha)^{\delta/2}\Gamma(\delta/2)} x^{\frac{\delta}{2}-1}\exp(-\frac{x}{2\alpha}) \tag{5}$$

$\forall x > 0$, where Γ is the gamma function. Considering that $\delta = 2$ and assuming that $\alpha = \frac{\gamma_n[n,k]}{2}$ and $\theta = 0$ for 'noise only' pixels, the pdf of H_0 is:

$$p_{H_0}(x) = \frac{1}{\gamma_n}\exp(-\frac{x}{\gamma_n[n,k]}) \tag{6}$$

The probability density function (pdf) of the spectrogram may deviate from a χ^2 law if real and imaginary part of $S_x^\omega[n,k]$ have different variances and if they are not independent variables. The use of a null boundaries window reduces mismatches between spectrogram and χ^2 pdfs [4].

2.2 Noise level estimation

2.2.1 Estimation via low-pass filtering

The smoothing of spectrogram coefficient is a widely used technique for NLE. A 1^{st} order infinite impulse response (IIR) low-pass filter applied to each frequency channel realizes a NLE via time recursive smoothing :

$$\hat{\gamma}_n[n,k] = (1+A)S_x^\omega[n,k] - A \times \hat{\gamma}_n[n-1,k] \tag{7}$$

with

$$A = \exp\frac{-(1-\tau_S)M}{\tau_i f_s} \tag{8}$$

where τ_S is the overlap ratio of $\omega[n]$, f_s the signal sampling frequency and τ_i the integration time of the smoothing filter. The major drawback of this technique is the constant integration time throughout a signal when the signal is composed of a wild variety of impulsive and frequency modulated sounds is not suited for the purpose of the detection scheme for the reasons mentioned in the introduction.

2.2.2 Estimation via statistical estimation

The principle of the statistical NLE proposed here is to extract the parameters the distribution of 'noise only' pixels in a time-frequency neighborhood of the current pixel. The NLE is less sensitive to abrupt change in the signal energy by taking into account not only the past pixels of the same frequency channel but the local time-frequency behavior of the current pixel. Given a mixture of noise and signal, we can make the assumption that the lowest energetic pixels are 'noise only' pixels. The statistical NLE is based on the distribution function of the lowest energetic pixels of a neighborhood. Let be $x_{(1)} \dots x_{(N)}$ a sorted set of N χ^2 variables with two degrees of freedom, the distribution function $F_X(x) = P[X < x]$ is:

$$F_X(x) = 1 - \exp(-\frac{x}{\gamma_n}) \tag{9}$$

then

$$\gamma_n = \frac{x}{-\ln(1-F_X(x))} \tag{10}$$

An empirical estimation of the distribution function is:

$$\hat{F_X}(x_{(i)}) = \frac{i}{N} \tag{11}$$

Given this estimation, a NLE for each of the N χ^2 variables $x_{(1)} \dots x_{(N)}$ is:

$$\hat{\gamma}_n = \frac{x_{(i)}}{-\ln(1-\frac{i}{N})} \tag{12}$$

Let be Z the number of noise only pixels in a neighborhood of N pixels, then the NLE chosen here is the one that has the lowest variance [5]:

$$\hat{\gamma_n} = \frac{x_{(Z)}}{-\ln(1-\frac{Z}{N})} \tag{13}$$

2.2.3 Evaluation of the statistical NLE

This section is dedicated to the evaluation of the statistical NLE in terms of its normalized bias and variance. The signal used here is a centered gaussian noise with a duration of 4 seconds and a sampling frequency of 48 kHz. The spectrogram is computed using a -180 dB Kaiser window of length 256 samples with an overlapping ratio of 0.5 and STFT length of 1024. The NLE is evaluated for only one pixel in the middle of the spectrogram and the evaluation is repeated 1000 times ($R = 1000$). The target value is the arithmetic mean of the pixel inside a time-frequency neighborhood of size 1001×1001 ($N = 10^6$):

$$\gamma_n = \frac{1}{R \times N} \sum_{j=1}^{R} \sum_{i=1}^{N} \gamma_{ji} \tag{14}$$

The parameters of the evaluation are the number of pixels in the neighborhood and the number of pixels considered as "noise only". A square-shaped neighborhood is used in the evaluation. Changing the shape would not bring additional information as the signal is stationary and white. The normalized bias and variance are computed as follow:

$$B_{\gamma_n} = \frac{E[\hat{\gamma_n}] - \gamma_n}{\gamma_n} \tag{15}$$

$$Var_{\gamma_n} = \frac{Var(\hat{\gamma_n})}{\gamma_n^2} \tag{16}$$

The figure Fig.1 depicts the normalized bias and log-variance as a function of the number of pixels in the neighborhood for various values of Z. Bias and variance are decreasing functions of number of pixels and of Z. As the signal contains no information but the noise, the noise estimation gets closer to the target value as the number of pixels used for the statistical estimation increases. For Z values equal to $\frac{N}{10}$ and higher, the asymptotic bias value is reached even with small neighborhood whereas for Z values lower than $\frac{N}{10}$ the neighborhood must have at least 10^4 pixels.

3 Results

3.1 Synthetic data

The statistical NLE is now applied to a simulated data to highlight the benefit of the method in comparison with the IIR smoothing filter. The sound file under study is made of a Gaussian white noise and three signals to be detected: one monochromatic frequency line, a wide-band transient to simulate a bio-acoustic click and an up-sweep between 2 kHz and 8 kHz to synthesize a whistle with lesser energy than the click. The comparison between both methods is made by plotting the detection maps. $\hat{\gamma}_n$ are computed for each pixel of the spectrogram and the pixel is detected if its energy is higher than 10 times the noise level (i.e. we consider $\lambda = 10$ dB). The parameters of the statistical NLE are a neighborhood of size $N = 21 * 21$ and $Z = 3N/4$ and the IIR integration time is 1 second.

The figure Fig.2 shows in the upper panel the spectrogram of the sound file (the frequency line cannot be seen because of the picture resolution), in the middle panel the detection map for the IIR method and in the bottom panel the detection map for the statistical NLE. Detected pixel are depicted with red squares. The IIR NLE shows clearly a masking effect due to the strong click and the integration time. After the click, no more pixels are detected. The frequency line is not detected at all. As the IIR NLE is computed for each frequency channel independently, the frequency line is considered as noise and no pixel with lesser energy could be detected on the corresponding frequency channel. The detection map of the statistical NLE shows all the three signals to be detected. No masking effect is observed and the frequency line is well defined thanks to use of a time-frequency neighborhood.

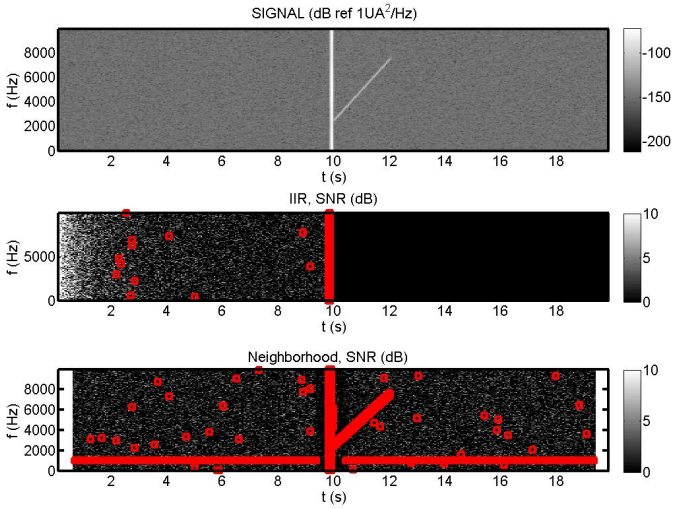

Figure 2: Detection maps for sperm-whales clicks and $\lambda = 10$ dB. Top panel, signal spectrogram with level in db ref 1 UA^2/Hz; middle panel, IIR NLE detection results with $\tau_i = 1$ s; bottom panel, statistical NLE detection results with neighborhood of size 21*21, $Z = 3N/4$.

3.2 Field data

The statistical NLE is now applied to field data. The sound file is a 16 bits resolution, 96 kHz frequency sampling recording from the NEMO observatory in Catane, Sicilia at 2000 m depth. The four panels of the figure Fig.3 depict four replications of the results of the statistical NLE with four neighborhood sizes : the top panel is for a 3*3 neighborhood, the middle-top panel is for a 5*5 neighborhood, the top middle-bottom panel is for a 21*21 neighborhood and the bottom panel is for a 51*51 neighborhood. The background noise is nearly stationary and but not white. We can see on the figure Fig.3 that the small neighborhoods overestimate the noise level and they are too sensitive to strong signal energy fluctuations. The presence of an impulse sound would lead to a blind period of detection. In the opposite, the use of large neighborhoods reduce the estimation bias and result in a smoothed NLE with a lower dynamic.

The figure in top panel of the Fig.4 shows the spec-

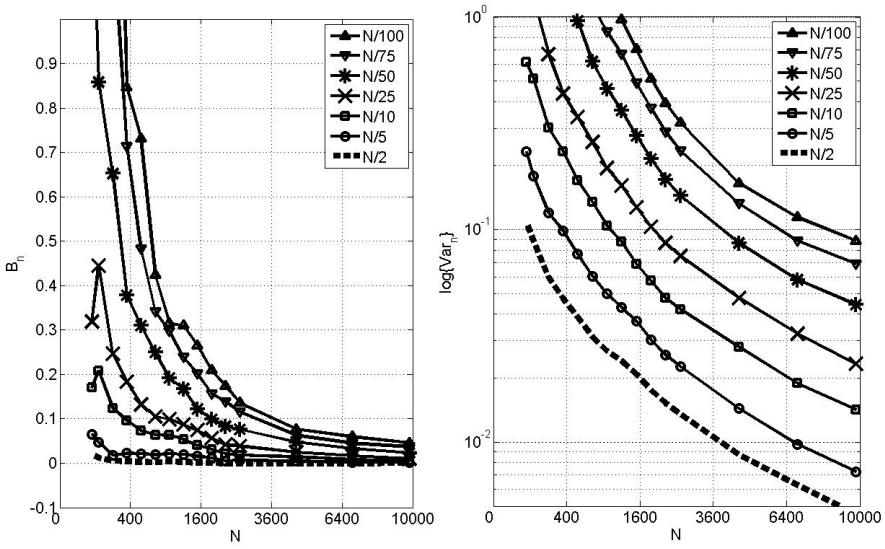

Figure 1: Normalized bias and log-variance for noise only signal as functions of the number of pixels in the time frequency neighborhood for various values of Z.

trogram of the recording. The vertical lines correspond to sperm-whales clicks. Whistles of risso dolphins are also present (cf. between 12 and 13 seconds). The figure in the middle panel of the Fig.4 depicts the detection map of the IIR NLE with $\lambda = 10$ dB and $\tau_i = 1$ s. The masking effect due to the clicks is visible and it leads to click misses. When two clicks are close in time the second is missed. The whistle between 12 and 13 second is not detected. The figure in the bottom panel of the Fig.4 shows the results of the detection via statistical NLE with a time-frequency neighborhood of size 21*21, $Z = N/2$ and $\lambda = 10$ dB. All the clicks are detected and the whistle between 12 and 13 second is well-defined even if a click appears simultaneously. The detection map of the statistical NLE show far more above threshold pixels and more energetic pixels (pixels with brighter gray level).

3.3 Computational time

The major drawback of the statistical MLE is its computation cost. For example, the statistical method requires 441 uses of the equation 14 for one NLE whereas the IIR method requires only 2 calculi (cf. eq (7)) for one NLE. The use of statistical methods could lead to implementation troubles when dealing with embedded hardware and real-time processing where both computational resources and processing time are critical. To tackle these problems, experiments involving interpolation and decimation are run to reduce required computational resources. The statistical NLE is evaluated only for the half the spectrogram pixels and this coarse NLE map is then refined thanks to a linear interpolation to the fit the original time-frequency resolution. The table Tab.1 shows measures of computational speed for a spectrogram of 9.10^5 pixels. The spectrogram is obtained for a 10 seconds signal with a sampling frequency of 44100 Hz, a FFT window size of 2048 and an overlap factor of 0.75. The computation are processed via an Intel Core 2 Duo computer with a 3.16 GHz clock and 3.35 Go of RAM. When the coarse computation grid is used, the computation speed increases dramatically and the computation time is below the real-time (1.3 s for a 10 s signal, spectrogram computation not taken into consideration). The table Tab.1 indicates results for two sizes of square-shaped neighborhood. The computation time reduces from 21.5 s to 7.5 s when neighborhood size decreases from 21*21 to 11*11. The last two entry of Tab.1 show that the computational time does not decrease when using a coarse computation grid and as the neighborhood size varies from 21*21 to 11*11. This is due to the increase of the number of pixels we account for when reducing the size of neighborhood in conjunction with keeping the coarse map resolution 50% lower than the original.

Table 1: Computational time of IIR and statistical NLE.

Method	Neighborhood size	Interpolation	Time (s)
IIR	non	non	0.008
statistical	21*21	non	21
statistical	11*11	non	7.5
statistical	21*21	oui	1.3
statistical	11*11	oui	1.3

4 Discussion

The results section substantiates that the statistical NLE enables far more pixels to be detected using a detection scheme based on the energy ratio between 'noise only' and 'noise + signal' pixels. The SNR threshold value can be lowered in comparison with IIR method. A decrease of the SNR threshold can be transposed into an increase in the detection range using the sonar equation :

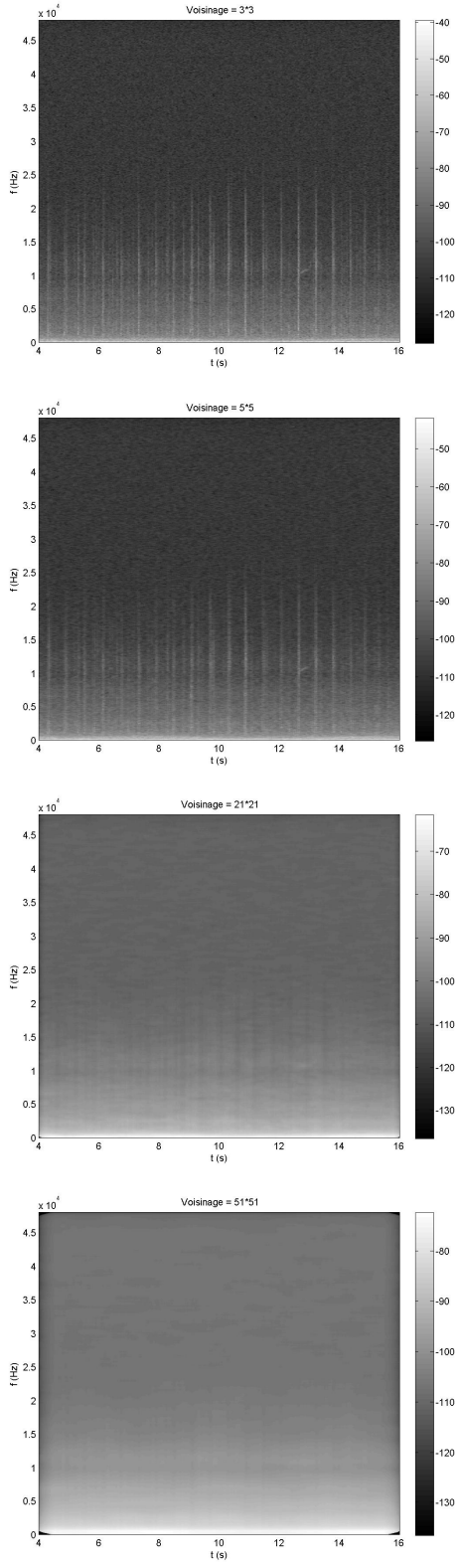

Figure 3: Spectrograms of statistical noise level estimation. Top panel: 3*3 neighborhood, middle-top panel: 5*5 neighborhood, middle-bottom panel: 21*21 neighborhood, bottom panel: 51*51 neighborhood.

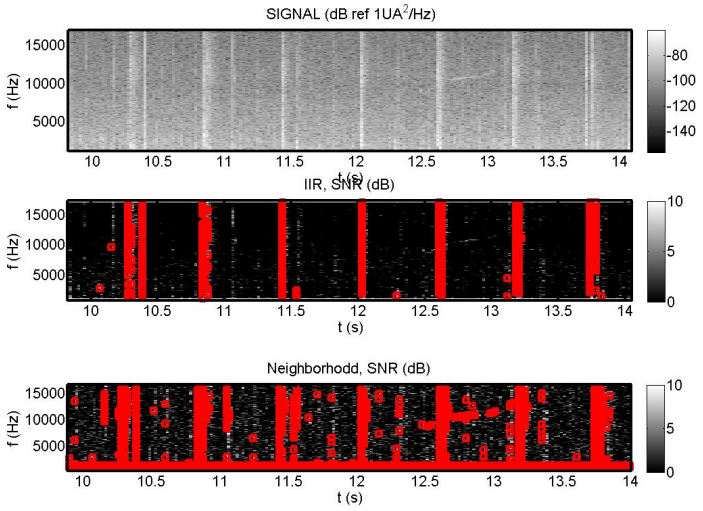

Figure 4: Detection map for sperm-whales clicks and $\lambda = 10dB$. Top panel, signal spectrogram with level in db ref 1 UA^2/Hz; middle panel, IIR NLE detection results with $\tau_i = 1s$; bottom panel, statistical NLE detection results with neighborhood of size 21*21, $Z = N/2$.

$$\lambda = SL - TL - \gamma_n \tag{17}$$

where SL is the sound level in db ref $1\mu Pa^2$ at 1 m, TL are the transmission losses and γ_n the noise level. If we consider spherical losses :

$$TL = 20 \log 10R \tag{18}$$

where is the detection range express in meter. It follows that lowering λ increases the range of detection:

$$R = 10^{\frac{SL-\lambda-\gamma_n}{20}} \tag{19}$$

As the number of detected pixel increase dramatically, the bio-acoustic stage of a PAM detector should be adapted to avoid an increase in the false alarm rate. The effort should be put in particular to avoid isolated pixels in the detection map.

The application of the statistical NLE to field data shows the interest of the method when multiple species are present, for example sperm-whales and dolphins, or when the species produce clicks and whistles, as dolphins do. Thanks to the statistical NLE, the whistles, which consist in frequency modulation tracks, are easier to extract from a raw recording. This is remarkably useful when dealing with density estimation purposes or when classification process required sharp edge detection. An other benefit of the statistical method is that it will helps tracking methods as Kalman filter [6] or time-frequency-phase analyzer [7] for the analysis of the whistles.

The major drawback of the method proposed here is its computational cost. Decimation and interpolation are successfully implemented to reach the real-time implementation. Depending of the environmental and recording conditions, the noise level could be either non stationary with highly fluctuating spectral features, as it

occurs in coastal area or in the vicinity of ship traffic, or stationary with fixed mean amplitude spectrum (cf. Fig4). If the noise environment is varying, the NLE should be done at high frequency in time and in frequency. In the opposite, the NLE could be estimated each one hour or at a lower frequency with a coarse resolution in frequency. The next step for upgrading the method is to estimate a level of stationarity and a level of whiteness of the signal prior to NLE. This would enable the method to adapt both the shape and the size of the neighborhood as well as the frequency and the time sampling. The computation cost would be dramatically reduced in case of stationary environment.

5 Conclusion

The estimation of the underwater noise level is a key point for the acoustic monitoring of undersea activities. One common estimation of the noise level is made by averaging the past signal level in each frequency channel of the spectrogram. Despite easy to implement, this methods suffers from being hard to tune to fast changing environment. The method proposed here takes into account both the time-frequency vicinity and the statistic properties of the current pixel neighborhood. The consideration of the behavior the pixel neighborhood leads to a 2D smoothing of the noise level. The estimation of the probability distribution function of the weaker energetic pixels enables a low bias and variance noise level estimation. Decimation and interpolation methods are useful to reduce the computation time at a level compatible with the real-time processing. One perspective of this work will be a study of the benefit of the interpolation method with respect to the quality of the NLE. Further works will involve a study a the stationarity and the whiteness of the underwater noise to automatically adapt the shape and the size of the neighborhood as well as the frequency and time estimation sampling.

Acknowledgments

The authors are thankful to Dr. Michel André from Laboratori d'Aplicacions Bioacústiques, Centre Tecnológic de Vilanova i la Geltrú, Universitat Politècnica de Catalunya, for providing sperm-whales sound.

References

[1] N.R.C., *Ocean noise and marine mammals*, (2003). The National Academies Press. Washington, D.C.

[2] D.K. Mellinger, K.M. Stafford, S.E. Moore , R.P. Dziak, and H. Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans", *Oceanography*, 20(4), (2007), 36-45.

[3] R. Hogg, and J.Ledolter, "Applied statistics for engineers and physical scientist", Second edition, *Macmilla Publishing company*, (1992), ISBN : 0-02-946409-9.

[4] J. Huillery, F. Millioz, and N. Martin, "On the Description of Spectrogram Probabilities with a Chi-Squared Law", *IEEE Transactions on Signal Processing*, 56(6), (2008), 2249-2258

[5] J. Huillery, "Support temps-fréquence d'un signal inconnu en présence de bruit additif gaussien", *PHD thesis*, 09 Juillet 2008, Institut polytechnique de Grenoble, France.

[6] A. Mallawaarachchi, S. H. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles", *The Journal of the Acoustical Society of America*, 124(2), 1159-1170.

[7] C. Ioana, and A. Quinquis, "Time-Frequency Analysis using Warped-Based High-Order Phase Modeling", *EURASIP Journal of Applied Signal Processing*, 2005(17), 2856-2873, Sept. 2005.