Audio Engineering Society

Convention Paper

Presented at the 115th Convention
2003 October 10–13 New York, NY, USA

A Sound Localizer Robust to Reverberation

José Vieira[1], and Luís Almeida[1]

[1]*Dep. de Electrónica da Universidade de Aveiro / IEETA, Aveiro, 3810-193, Portugal*

Correspondence should be addressed to José Vieira (vieira@det.ua.pt)

ABSTRACT

This paper proposes an intelligent acoustic sensor able to localize sound sources in acoustic environments with strong reverberation. The proposed system is inspired on the precedence effect used by the human auditory system and uses only two acoustic sensors. It implements a modified version of the algorithm proposed by Huang that uses the precedence effect in order to achieve robust sound localization even in reverberant environments. The localization system was impletented in a C31 DSP for real time demonstration and several experiments were performed showing the validity of our solution. Finally, the paper also proposes a method to estimate on-line the decay of the reverberation, using the received sound signals, only.

1. INTRODUCTION

The ability of our auditory system to localize sounds has a great importance in our everyday life. Developing an artificial system, able to localize sounds in the same way as humans do could also bring along many advantages in diverse application fields such as video conferencing, video surveillance, intelligent homes and robotics. It is, however, a very challenge

task[1].

As sound localization is a feature of humans' auditory system (and of most mammals, too), it has been deeply studied in the fields of physiology and psychoacoustics [2–5]. From the substantial amount of literature available, one can find a vast range of materials including experimental work on binaural perception, but without attempting to find a computational model for the observed phenomena, or, as in [2–4, 6], signal processing models that explain or mimic the binaural human auditory system. In a parallel direction, and most of the times without cross references, one can also find a substantial amount of literature in the field of "smart sensors" and "microphone array signal processing" that present sophisticated algorithms to solve the problems of sound localization without any inspiration on the human auditory system (see [7] for a up-to-date review).

In this paper we present a system able to determine the direction of arrival of any type of sound with only two microphones, using the Interaural Time Difference (ITD). Along the text and in sake of simplicity, we will refer to the determination of such direction of arrival as sound localization. One problem that greatly affects the localization accuracy is the reverberation noise. Therefore, to overcome such problem we used an artificial model of the well known precedence effect. This method uses an approximate model of the room acoustics to predict, based on the received signal, the amount of reverberation noise and the instants, called onsets, where the direct signal to reverberation noise ratio is favorable. An interesting feature that is now subject of on-going work is the on-line estimation of the room acoustics so that the previous method can self-adapt to its operating environment.

This article is organized as follows, section 2 presents a brief introduction to the relevant aspects of the human auditory system that inspired our work. The following two sections present, respectively, the method used for sound localization based on the ITD and the method to compute it based on the cross correlation of the signals received at both left and right microphones. Section 5 describes the model of the precedence effect and section 6 shows several experimental results of the proposed algorithm, executed on a TMS320C31-based platform. Finally, section 7 presents some preliminary work towards the on-line estimation of the acoustic parameters of the room used in our precedence effect model and section 8 concludes the paper.

[1] For a interesting introduction to the human hability to localize sounds see [1]

2. SOUND LOCALIZATION

The ability to localize sounds is of paramount importance to most of the mammals. Several studies about the animal evolution revealed that the audible frequency range is related with the necessity of sound localization, and the acuity of the sound localization system is related to the dimension of the central vision field, which has more resolution [5]. Moreover, a fast and correct detection of a sound source is a great competitive advantage for the survival of the species.

The human auditory system is able to detect the azimuth and the elevation of a sound source, even if the sound is behind the listener, and not loosing acuity in reverberant acoustic environments [1, 2]. Despite having two acoustic sensors, only, this remarkable skill is achieved by the auditory system with an elaborate processing of the acoustic signals.

For the determination of the azimuth of a sound source the human auditory system uses two different techniques [2]: The Interaural Level Difference (ILD), and the Interaural Time Difference (ITD). The first method is applied to localize sounds with a wavelength shorter than the diameter of the head (about 2kHz) and it exploits the shadow effect that the head causes at the ear located in the opposed side with respect to the sound source. The second method, for sounds with a wavelength longer than the diameter of the head, relies on the measurement of the relative delay of a sound arriving at the two ears.

None of these techniques explain the ability of the human auditory system to localize sounds in reverberant environments. This capacity is achieved by the precedence effect also known as the "law of the first wave front", which was discovered experimentally by the Princeton physician Joseph Henry (1797-1878) [8]. Henry verified that when two similar sounds arrive to a listener in sequence and within

a small time interval, only the origin of the first one is perceived, even if they were originated at two different locations.

3. SOUND LOCALIZATION USING THE ITD

Consider a system to localize a sound source using only two sensors as described in figure 1. If the distance to the sound source is sufficiently larger than the distance d between the two microphones, we can consider that the received acoustic waves are planar.

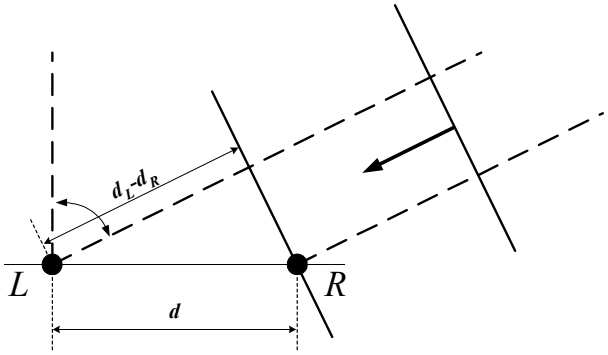

Fig. 1: Geometry of the sound localization problem. The two microphones located at points L and R receive a planar wave front coming from the right.

If the sound source is localized at distance d_L and d_R from the left and right microphones respectively, then, from the time difference Δt of the wave front arriving at the microphones, we can evaluate the angle θ of the sound source (1), with v_c the sound speed.

$$\sin(\theta) = \frac{d_L - d_R}{d} = \frac{v_c \Delta t}{d}. \qquad (1)$$

4. MEASURING THE ITD USING CROSS CORRELATION

A simple model for the measured signals in the left and right microphones is given by

$$\begin{aligned} x_L(t) &= s(t) + n_L(t) \\ x_R(t) &= s(t - \Delta t) + n_R(t), \end{aligned}$$

where $s(t)$ is any sound source signal and $n_L(t)$ and $n_R(t)$ are uncorrelated white noise generated by the microphones. In our system these two signals are low pass filtered and sampled at 8kHz leading two the following sampled versions

$$\begin{aligned} x_L(n) &= s(n) + n_L(n) \\ x_R(n) &= s(n - \Delta t) + n_R(n). \end{aligned}$$

Notice the fractional time delay Δt that we must estimate. To evaluate the time delay between the signal $x_L(n)$ and $x_R(n)$ we can perform the cross-correlation between these signals and find the maximum. The signals cross correlation is obtained by the expression

$$R(k) = \sum_{n=0}^{K-1} x_L(n) x_R(n+k) \qquad k \in [-N \ldots + N],$$

where N is the number of samples corresponding to maximum lag of the correlation. The maximum value for Δt is given by

$$\Delta t_{\max} \leq \frac{d}{v_c}$$

leading to

$$N = \left\lceil \frac{f_s d}{v_c} \right\rceil,$$

where $\lceil \cdot \rceil$ (ceiling function) means the rounding operation to the next integer.

For a sampling frequency f_s of 8000Hz, and a distance between the two microphones of 0.33 meters, we have a limited resolution with the following angles: $0^o, \pm 8^o, \pm 16^o, \pm 24^o, \pm 33^o, \pm 42^o, \pm 54^o, \pm 70^o, \pm 90^o$. To improve the resolution of this method we can increase the sampling rate of the input signals. However, with this solution the number of arithmetic operations increases as $\mathcal{O}(f_s^2)$. It would be computationally simpler to interpolate the samples of the cross-correlation to increase the resolution, while maintaining the sampling frequency. However, for this approach to be valid, it must be demonstrated that the frequency contents of the cross-correlation are limited and, particularly, limited to the bandwidth of the original signal. Taken this statement as correct it is possible, for example, to use cubic splines to perform the interpolation of the cross-correlation points and obtain the desired resolution (figure 2).

Therefore, we state and prove the following lemma:

Lemma 1 *Consider a signal $u(t)$, and $x(t)$ a lowpass version, filtered by an ideal lowpass filter with frequency response*

$$H(f) = \begin{cases} 1 & |f| < f_s/2 \\ 0 & |f| \geq f_s/2 \end{cases}.$$

Then, the cross-correlation of $x(t)$ with $x(t-\Delta t)$ is also bandlimited to $f_s/2$.

Proof. The auto-correlation of the signal $x(t)$ can be written as a function of the auto-correlation of the signal $u(t)$

$$R_x(\tau) = h^*(-\tau) * h(\tau) * R_u(\tau),$$

and if we define $S_u(f)$ and $S_x(f)$ as the Fourier transform of $R_u(\tau)$ and $R_x(\tau)$ respectively, we have

$$S_x(f) = |H(f)|^2 S_u(f).$$

From this equation we can see that $R_x(\tau)$ is a band limited signal, having the same bandwidth as $x(t)$, and we can use an interpolation technique to obtain values of $R_x(\tau)$ from the sampled version $R_x(k)$. Now considering our case where $x_L(t) = x(t)$ and $x_R = x(t-\Delta t)$, the evaluation of the cross-correlation between the left and right signals is given by

$$R_{x_L x_R}(\tau) = \int_{-\infty}^{+\infty} x_L(t) x_R(t-\tau) dt =$$
$$= \int_{-\infty}^{+\infty} x(t) x(t-\Delta t-\tau) dt = R_x(\tau+\Delta t),$$

which is just a delayed version of $R_x(\tau)$ and for this reason is also a band limited signal with the same bandwidth as $x(t)$. ■

One alternative way to compute the ITD not based on time correlation is to multiply the cross spectrum of each signal [9]. Consider the DFT of the signals x_L and x_R leading to the spectrums $X_L(e^{j\omega})$ and $X_R(e^{j\omega})$, respectively. After evaluating the product $X_L(e^{j\omega})X_R^*(e^{j\omega})$, where the symbol * denotes the complex conjugate, we can calculate the inverse discrete Fourier transform of the result, obtaining the cross-correlation signal. A simple peak detection can be used to estimate the time delay Δt. For our application, this method, despite the use of the FFT algorithm to evaluate the cross-correlation, is inefficient because we only use 17 samples of the result, and in the case of the time domain cross-correlation we only evaluate the necessary samples of $R_{x_L x_R}(\tau)$. Moreover, these two methods are equivalent and give the same results.

The method proposed by Huang [10] uses yet another possibility to estimate the time delay (ITD), which is to evaluate the DFT of the signals $x_L(n)$ and $x_R(n)$ and measure the phase difference between the DFT components for the low frequency components to avoid phase wrapping. Despite its lower computational cost, we believe that this method is less accurate than the one proposed in this paper. The data available in the literature does not allow a direct comparison but we expect to carry out comparative experiments in near future work.

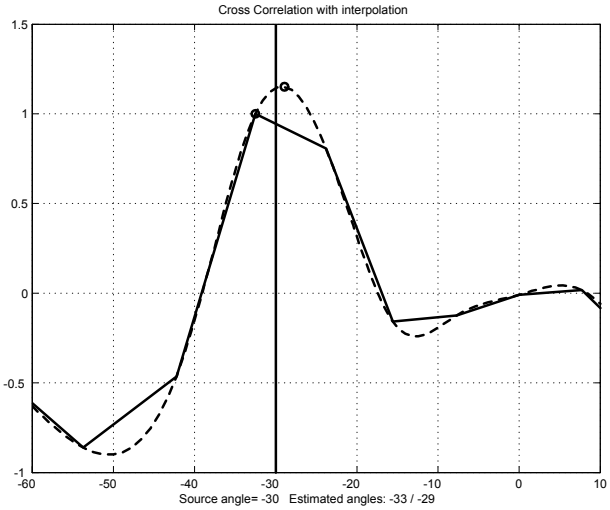

Fig. 2: Interpolation of the discrete cross-correlation to increase the accuracy of the angle estimation algorithm. The source angle was -30^o.

5. MODELING THE PRECEDENCE EFFECT

The methods presented in the previous section to determine the direction of a sound source fail when used in reverberant environments. This problem is of great relevance because reverberation is present in most of the foreseen operational environments, such as rooms in buildings. This problem, however, can be overcome using the precedence effect, which has

been deeply studied since a long time [3, 4, 11–16]. Nevertheless, despite all those studies, the only references to artificial models of the precedence effect are those from Huang etal [6, 10, 17, 18]. This is a computational model based on the psychoacoustics study of that effect, according to which there is an inhibition of the sound localization capacity after the arrival of an acoustic wave front. This phenomenon allows the localization system to avoid the use of signal contaminated with reverberation since it discards the "echoes" that might follow the arrival of an upcoming wave front. The detection of wave fronts is based on an adaptive sound level threshold. If a wave front is followed by another one of greater amplitude, the inhibition is canceled and the localization is turned on again, for the newly arrived wave front. The model proposed by Huang etal is based on a dynamic modeling of the amount of reverberation that is detected in the operational environment, taking into account previous sounds emitted in the same place.

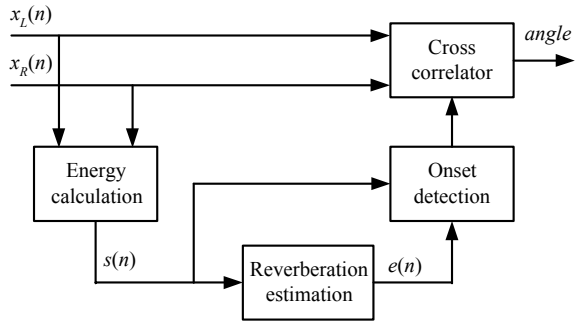

Fig. 3: The system measures the energy of the inputs and estimates the reverberation noise $e(n)$. The onset detector compares the amplitude of the signal $s(n)$ with $e(n)$ and performs a cross-correlation on the onsets.

Our version of this model can be observed in the block diagram of figure 3. We start by evaluating the mean energy of both channels (signal $s(t)$), by taking the mean of the module of the input signals and filtering the result using a first order IIR low pass filter. This signal is used as input to a reverberation estimator of the environment. The estimation of the reverberation $e(n)$ is constantly compared to the input energy $s(n)$, and when the following condition is verified

$$\frac{s(n)}{e(n)} > threshold,$$

the system considers that it is in the presence of an "onset" and performs a cross-correlation.

The model for the reverberation estimator, assumes that the energy of the reverberation decreases exponentially. The proposed impulse response is given by equation (2). In this model, there is a time delay T between the wave front arrival instant and the beginning of the reverberation. The portion of sound signal captured during this interval is considered free from reverberation and it is used for the cross correlation between the received signals at both microphones. After this interval, the reverberation level decays exponentially with time constant τ. The constant g gives the attenuation between the wave front and the first reflection.

$$h_p(t) = \begin{cases} 0 & 0 \leq t \leq T \\ ge^{(t-T)/\tau} & t > T \end{cases} \quad (2)$$

In figure 4 we can see the response of this system to four impulses of different amplitudes. The solid line is the estimated echo signal $e(n)$. The received impulses 1 and 3 are good candidates for onsets, because they have a favorable signal to reverberation noise ratio. Conversely, the impulses 2 and 4 have an amplitude below the reverberation noise level. In section 6 some examples with real signal will be presented.

6. EXPERIMENTAL RESULTS

In order to assess the performance of the sound localizer, several sound signals were acquired in 3 different rooms with different dimensions and diverse acoustic responses. Different signals were also recorded in order to test the robustness of the algorithm, namely a hands clap, the vowel "a", the "psst" sound and the sentence "sound experiment" (the sounds are available in http://www.ieeta.pt/~vieira/proj5ano/2001-02/goodears/sons.html). All these sounds were recorded for several angles (-60^o,-30^o,0^o,30^o,60^o). Despite possibles differences in the acoustics of the three rooms, the following set of reverberation parameters was used for all of them: $\tau = 214ms$,

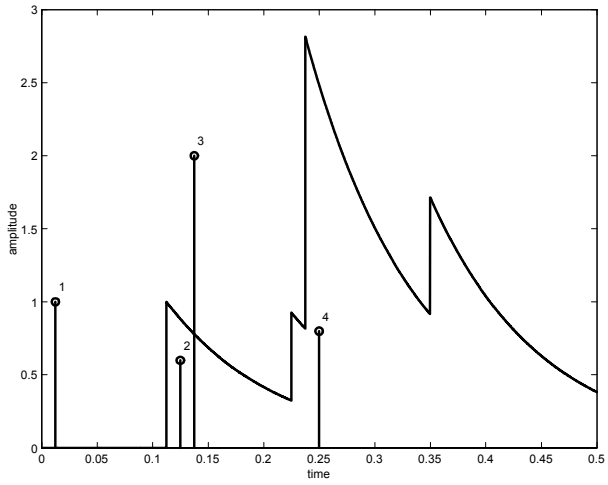

Fig. 4: Response of the reverberation model to the impulses 1 to 4. The estimated reverberation noise is the solid line.

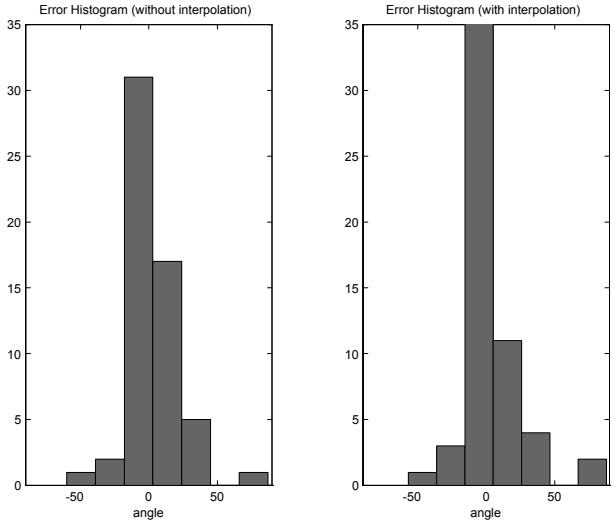

Fig. 5: Left: error histogram using direct correlation between left and right channels with peak detection. Right: error histogram using interpolation between the cross-correlation samples to increase accuracy.

$T = 9ms$ and $g = 0.5$. Figure 5 shows two histograms of the error in estimating the source angle. The left histogram was generated with the cross-correlation algorithm without interpolation and the one on the right was obtained using interpolation. Notice that the correct angle is detected with a reasonable precision despite the diversity of acoustic signals and rooms used. We can also see the increase in the precision of angle estimation with the interpolation.

To verify the negative impact of reverberation, we performed the cross-correlations at 200ms after each onset, well within the respective interference area. The error histogram of the sound localization is depicted in figure 6 where the degradation of the performance of the algorithm is clear.

7. AUTOMATIC EXTRACTION OF THE ACOUSTIC PARAMETERS

One of the limitations of the method presented before is the need to use predetermined parameters to estimate the reverberation noise of each room, i.e. T, τ and g. This can be done estimating the impulse response of the room acoustics, either measuring it directly with an impulse excitation, or indirectly using white noise and cross-correlating the original and received signals [19]. Then, from the energy of the impulse response it is possible to obtain the decay. Furthermore, when the source signal is unknown but some statistical model is available, one can also use a "blind" algorithm to estimate the impulse response of the room acoustics or, which is easier, just the decay [20].

Nevertheless, it would be desirable to estimate such parameters on-line, using the received sound signals, only. This would make the method self-adapting to the room and thus, more flexible. Therefore, in this section we show some preliminary work in this direction, with a simple method to estimate on-line the reverberation decay τ. This method does not rely on test signals (impulses or noise) and allows localizing any kind of sounds: voice, hand claps, etc, thus not assuming any statistical knowledge of the received signal.

Consider the signal $X(n)$ obtained from the received signal $x(n)$ using equation (2)

$$X(n) = 10 \log_{10} \left(\frac{1}{N} \sum_{k=n}^{n+N-1} x^2(k) \right),$$

where the constant N is the number of samples of the FIR low-pass comb filter. The signal $X(n)$ is the

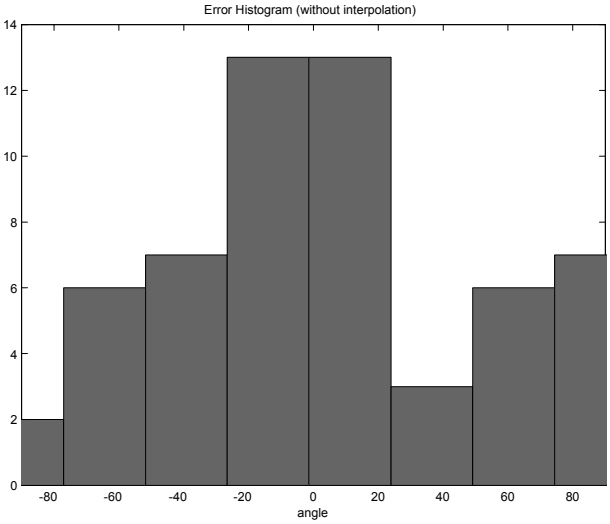

Fig. 6: Error histogram obtained using data 200ms after the onsets. The performance of the estimator was degraded due to the reverberation noise.

logarithm of the filtered energy of the signal $x(n)$, and the decay of the reverberation is transformed from exponential to linear. Figure 7 depicts the log-energy of two different recordings with quite different decays.

The slopes of the peak "tails" in each plot allow determining the decay parameter τ. For the sake of simplicity, the preliminary version of our algorithm assumes that the noise floor is known so that a threshold can be statically defined above it. However, an adaptive threshold can easily be implemented, for example using a similar technique to that of the onsets detection. Basically, the algorithm computes $X(n)$ from the original signal $x(n)$. Then, it scans $X(n)$ looking for cross-overs with respect to the threshold referred above. It generates a set of data points in the vicinity of each of those cross-overs, only if all points are monotonically decreasing, i.e. negative slopes (otherwise the set is discarded). These sets are then subject to a linear regression to estimate the decay τ and the T_{60}[2] parameter using

[2] The T_{60} parameter is defined as the time that the impulse response of a room takes to decreases 60dB.

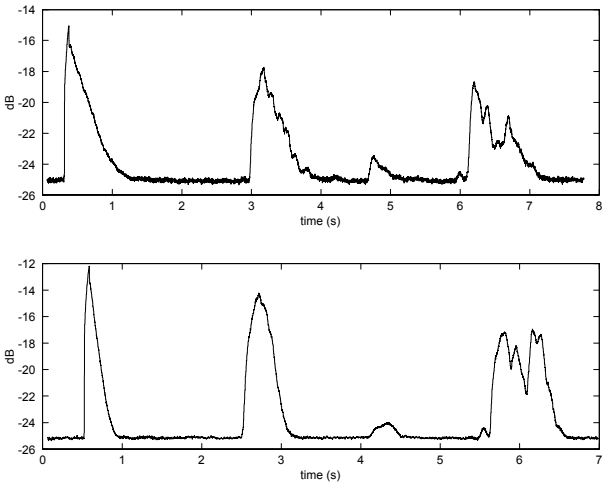

Fig. 7: Mean log-energy of two input signals (see text). The "tails" of the peaks allow determining the decay parameter.

the following equations

$$\tau = -\frac{10\log_{10} e}{\tau f_s} \qquad T_{60} = 6\log(10)\tau.$$

T_{60} is a common parameter in acoustics, meaning the time the reverberation energy takes to decrease 60dB from its initial value. Figure 8 shows an example of using this method, with a zoom on the first negative slope (below), where we can see the set of data points (circles) and the linear regression (straight line).

8. CONCLUSIONS

The main contribution of the paper is the actual construction of an artificial sound localization system that is capable of operating in the presence of strong reverberation using the precedence effect, as most mammals do. The localization algorithm is a modified version of the one proposed by Huang to localize sounds in reverberant environments. Several experimental results have been gathered that confirm the capacity of the system to correctly determine the sound source direction for very different sounds and in reverberating environments.

Future work will focus on the on-line estimation of the reverberation parameters using the actual received sound signals, only. This paper already in-

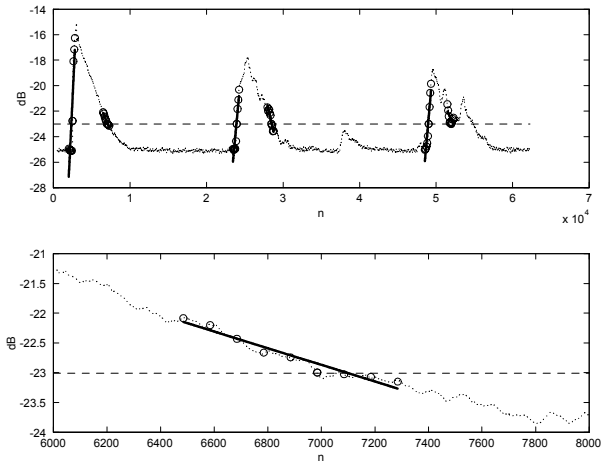

Fig. 8: In the top plot we can see the dashed horizontal line with the threshold level. The algorithm detected several useful points to measure the decay parameter τ. In the bottom plot we see a zoom of the first of the points with negative slope.

cludes a preliminary work in this direction, presenting a method to estimate on-line the reverberation decay. With such on-line characterization of the reverberating environment, the sound localizer can self-adapt to the room resulting in higher flexibility.

9. ACKNOWLEDGEMENTS

The authors acknowledge the help given by José Fonte and Daniel Gonçalves in an earlier stage of the work and, particularly, in the acquisition of the sounds used in the experiments.

10. REFERENCES

[1] William M. Hartmann, "How we localize sound," *Physics Today*, pp. 24–29, Nov. 1999.

[2] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Mit Press, 1996.

[3] W. Lindemann, "Extension of a binaural cross-correlation model by means of contralateral inhibition, i: Simulation od lateralization of stationary signals," *Journal of Acoustical Society of America*, vol. 80, no. 6, pp. 1602–1622, Dec. 1986.

[4] W. Lindemann, "Extension of a binaural cross-correlation model by means of contralateral inhibition, ii: The law of the first wave front," *Journal of Acoustical Society of America*, vol. 80, no. 6, pp. 1623–1630, Dec. 1986.

[5] R. S. Heffner and Henry E. Heffner, *The Evolutionary Biology of Hearing*, Springer-Verlag, 1992.

[6] Jie Huang, Noboru Ohnishi, and Noboru Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Transactions on Instrumentation and Measurement*, vol. 4, no. 4, pp. 842–846, Aug. 1997.

[7] M. Brandstein and D. (Eds.) Ward, *Microphone Arrays - Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

[8] Manfred R. Schroeder, *Computer Speech - Recognition, Compression, Synthesis*, Information Sciences. Springer, Berlin, 1999.

[9] Joseph C. Hassab and Ronald E. Boucher, "Optimum estimation of time delay by a generalized correlator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 373–380, Aug. 1979.

[10] Jie Huang, N. Ohnishi, X. Guo, and N. Sugie, "Echo avoidance in a computational model of the precedence effect," *Speech Communication (Elsevier Science)*, vol. 27, no. 3, pp. 223–233, 1999.

[11] Lloyd A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 61, pp. 468–486, 1948.

[12] Werner Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *Journal of Acoustical Society of America*, vol. 94, no. 1, pp. 98–110, July 1993.

[13] Markus Bodden, "Binaural modeling and auditory scene analysis," in *IEEE 1995 Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York, 1995, IEEE.

[14] C. Schauer, T. Zahn, P. Paschke, and H.-M. Gross, "Binaural sound localization in a artificial neural network," in *ICASSP 2000*, Instambul, Turkey, June 2000, IEEE, pp. 865–868.

[15] Paul Hofman and John van Opstal, "Identification of spectral features as sound localization cues in the external ear acoustics," in *Proceedings IWANN'97*, J.J. Mira-Mira, R. Moreno-Diaz, and J. Cabestany, Eds. 1997, pp. 1126–1135, Heidelberg: Springer.

[16] Paul M. Hofman and John Van Opstal, "Spectro-temporal factors in two-dimensional human sound localization," *Journal of Acoustical Society of America*, vol. 103, no. 5, pp. 2634–2648, May 1998.

[17] Jie Huang, Noboru Ohnishi, and Noboru Sugie, "Building ears for robots: Sound localization and separation," *Artificial Life and Robotics (Springer-Verlag)*, vol. 1, no. 4, pp. 157–163, 1997.

[18] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model based sound localization systems and its application to robot navigation," *Robotics and Autonomous Systems (Elsevier Science)*, vol. 27, no. 4, pp. 199–209, 1999.

[19] Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, and James V. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons. Inc., USA, 3 edition, 1982.

[20] Laurent Couvreur, Christophe Ris, and Couvreur Christophe, "Model-based blind estimation of reverberation time: Application to robust ASR in reverberant environments," in *EUROSPEECH-2001*, Aalborg, Denmark, Sept. 2001, vol. 4, pp. 2631–2634.