MASTER THESIS

TITLE: Retroactive Spatializer

MASTER DEGREE: Master in Science in Telecommunication Engineering & Management

AUTHOR: Lídia Regalado Sánchez

DIRECTOR: Sylvain Marchand

DATE: July 20, 2009

TABLE OF CONTENTS

INTRODUCTION

The aim of this project is increasing the research carried out by Sylvain Marchand, researcher of the LaBRI (Laboratoire Bordelais de Recherche en Informatique), about the RetroSpat System.

RetroSpat stands for "Retroactive Spatializer" and it is a free software project for the spatialization of sounds represented in the spectral domain. This system is intended to be a spatializer with perceptive feedback. More precisely, RetroSpat can guess positions of physical sound sources (e.g. loudspeakers) from binaural inputs, and can then output multichannel signals to the loudspeakers while controlling the spatial location of virtual sound sources.

So, the project will be based practically on improving the localization method as a result of the localization corresponding to a sound source, changing the software architecture to merge the localization and spatialization parts and on an improvement of the graphical user interface.

This project is organized as follows. In the first section is presented some history of *acousmatic music*. Next a description of the model in Section 2 is done including the spatialization and localization methods. Section 3 is dedicated to the Source code. The problems with the actual software and its possible reasons of appearance are discussed in Section 4. Finally there are some conclusions of the work done.

1 ACOUSMATIC MUSIC

A. History

Over centuries, the music has continuously undergone various innovations. In 1948, Schaeffer and Henry at the “Radio T´el´evision Franc¸aise” were interested in the expressive power of sounds. They used microphones to capture sounds, discs as supports, and transformation tools. The *musique concr`ete* was born.
In 1949, Eimer gave birth to *electronic music* in the studios of the German radio “Nordwestdeutscher Rundfunk” in Cologne. This music was produced by frequency generators. Koenig and Stockhausen were among the first to use it.
The merge of *musique concr`ete* and *electronic music* gave rise to *electro-acoustic music* or *acousmatic music*. Today, many musical pieces are created worldwide. Acousmatic has become a discipline that is taught in universities and conservatories.

B. Actual Practices

Composers of acousmatic music use both electronic and natural sounds recorded close to a microphone, such as wind noise, voices, wrinkling paper, etc. The sounds are then processed by a computer and organized by editing and mixing. The result is a *musical composition*.
However, the creation gets its full value when it is played in concert using an *acousmonium*: an orchestra of loudspeakers. The acousmonium consists of a highly variable number of loudspeakers with different characteristics. The interpreter of the piece controls the acousmonium from a special (un)mixing console.
The originality of such a device is to map the two stereo channels at the entrance to 8, 16, or even hundreds of channels of projection. Each channel is controlled individually by knobs and equalization systems. The channel is assigned to one or more loudspeakers positioned according to the acoustical environment and the artistic strategy.

C. Expected Improvements

Behind his/her console, the interpreter of acousmatic music acts in real time on various sound parameters such as spatial location, sound intensity, spectral color. He/She broadcasts a unique version of the music fixed on a medium. The acousmatic diffusion requires some skills.

RetroSpat intends to facilitate the work of the interpreter by improving the following embarrassing practices:

- two wheels needed to spatialize one source;
- stereo sources as inputs;
- no individual source path, only one global mix path;
- the distance spatialization requires some expertise.

2 MODEL

We consider a punctual and omni-directional sound source in the horizontal plane, located by its (ρ,θ) coordinates, where ρ is the distance of the source to the head center and θ is the azimuth angle. Indeed, as a first approximation in most musical situations, both the listeners and instrumentalists are standing on the (same) ground, with no relative elevation.
The source s will reach the left (L) and right (R) ears through different acoustic paths, characterizable with a pair of filters, which spectral versions are called Head-Related Transfer Functions (HRTFs). HRTFs are frequency- and subject-dependent. The CIPIC database [2] samples different listeners and directions of arrival.
A sound source positioned to the left will reach the left ear sooner than the right one, in the same manner the right level should be lower due to wave propagation and head shadowing. Thus, the difference in amplitude or Interaural Level Difference (ILD, expressed in decibels – dB) [4] and difference in arrival time or Interaural Time Difference (ITD, expressed in seconds) [5] are the main spatial cues for the human auditory system [6].

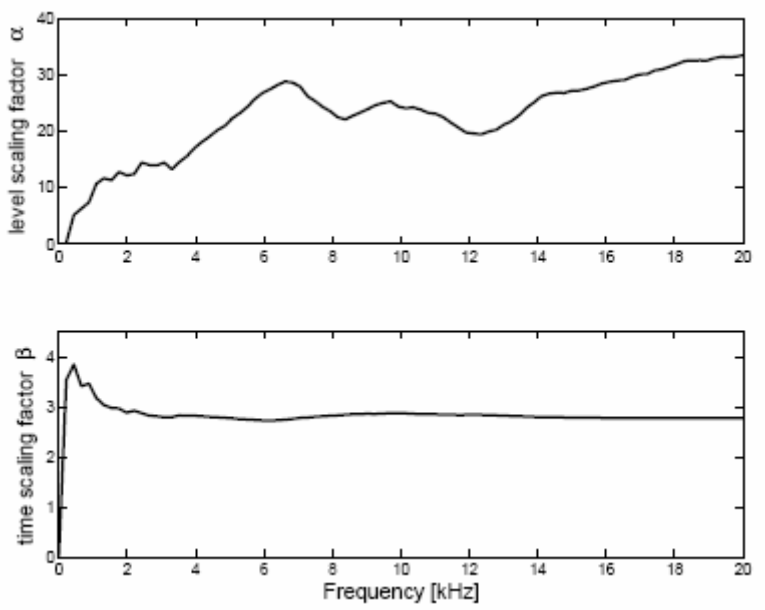


Fig. 1 Frequency-dependent scaling factors: α (top) and β (bottom)

A. Interaural Level Differences

After Viste [1], the ILDs can be expressed as functions of $\sin(\theta)$, thus leading to a sinusoidal model:

$$ILD(\theta,f)=\alpha(f)\sin(\theta) \quad (1)$$

where $\alpha(f)$ is the average scaling factor that best suits our model, in the least-square sense, for each listener of the CIPIC database (see Figure 1). The overall error of this model over the CIPIC database for all subjects, azimuths, and frequencies is of 4.29dB. The average model error and inter-subject variance are depicted in Figure 2.

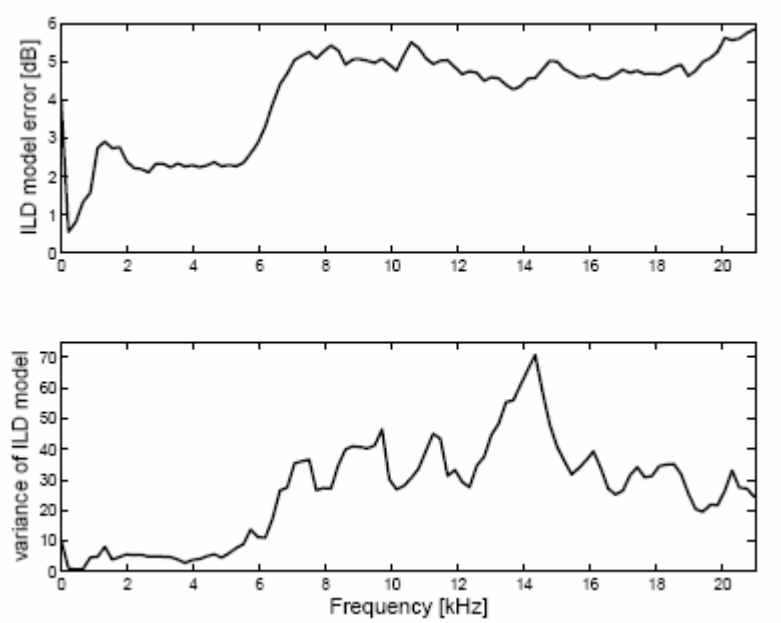


Fig. 2 Average ILD model error (top) and inter-subject variance (bottom)

Moreover, given the short-time spectra of the left X_L and right X_R channels, we can measure the ILD for each time-frequency bin with:

$$ILD(t,f) = 20\log_{10}\left|\frac{X_L(t,f)}{X_R(t,f)}\right| \quad (2)$$

B. Interaural Time Differences

Because of the head shadowing, Viste uses for the ITDs a model based on $\sin(\theta)+\theta$, after Woodworth [7]. However, from the theory of the diffraction of an harmonic plane wave by a sphere (the head), the ITDs should be proportional to $\sin(\theta)$. Contrary to the model by Kuhn [8], our model takes into account the inter-subject variation and the full-frequency band. The ITD model is then expressed as:

$$ITD(\theta,f) = \beta(f)\, r \sin(\theta)/c \quad (3)$$

where β is the average scaling factor that best suits our model, in the least-square sense, for each listener of the CIPIC database (see Figure 1), r denotes the head radius, and c is the sound celerity. The overall error of this model over the CIPIC database is 0.052ms (thus comparable to the 0.045ms

error of the model by Viste). The average model error and inter-subject variance are depicted in Figure 3.
Practically, our model is easily invertible, which is suitable for sound localization, contrary to the $\sin(\theta)+\theta$ model by Viste which introduced mathematical errors at the extreme azimuths (see [9]).
Given the short-time spectra of the left X_L and right X_R channels, we can measure the ITD for each timefrequency bin with:

$$ITD_p(t,f)=\frac{1}{2\pi f}\left(\angle\frac{X_L(t,f)}{X_R(t,f)}+2\pi p\right) \quad (4)$$

The coefficient p outlooks that the phase is determined up to a modulo 2π factor. In fact, the phase becomes ambiguous above 1500Hz, where the wavelength is shorter than the diameter of the head.

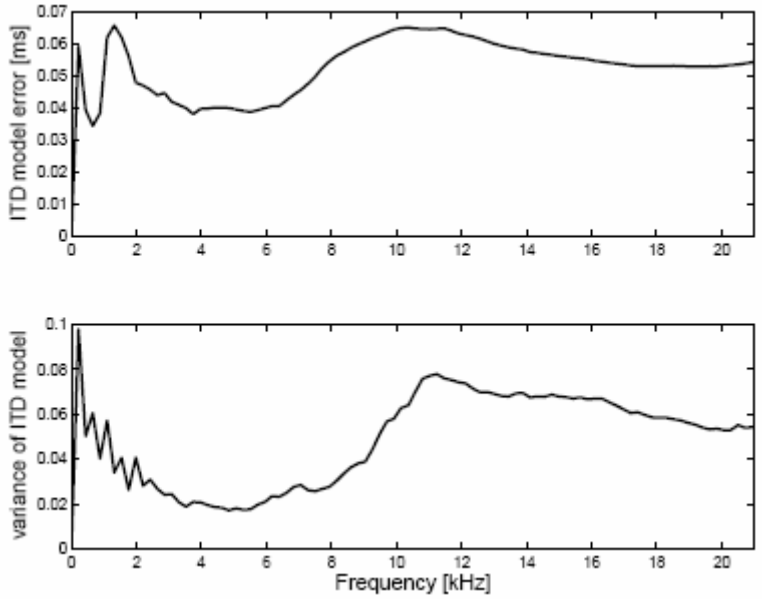


Fig. 3 Average ITD model error (top) and inert-subject variance

C. Distance Cues

The distance estimation or simulation is a complex task due to dependencies on source characteristics and the acoustical environment. Four principal cues are predominant in different situations: intensity, direct-to-reverberant (D/R) energy ratio, spectrum, and binaural differences (noticeable for distances less than 1m). Their combination is still an open research subject. Here, we focus effectively on the intensity and spectral cues.
In ideal conditions, the intensity of a source is halved (decreases by −6dB) when the distance is doubled, according to the well-known Inverse Square Law. Applying only this frequency-independent rule to a signal has no effect on the sound timbre. But when a source moves far from the listener, the high frequencies are more attenuated than the low frequencies. Thus the sound spectrum changes with the distance. More precisely, the spectral centroid

moves towards the low frequencies as the distance increases. In [10], the authors show that the frequency-dependent attenuation due to atmospheric attenuation is roughly proportional to f^2, similarly to the ISO 9613-1 norm. Here, we manipulate the magnitude spectrum to simulate the distance between the source and the listener (see Section IV). Conversely, we measure the spectral centroid (related to brightness) to estimate the source's distance to listener (see Section V).

2.1 Spatialization Method

A. Relative Distance Effect

In a concert room, the distance is often simulated by placing the speaker near / away from the auditorium, which is sometimes physically restricted in small rooms. In fact, the architecture of the room plays an important role and can lead to severe modifications in the interpretation of the piece.
Here, simulating the distance is a matter of changing the magnitude of each short-term spectrum X. More precisely, the ISO 9613-1 norm gives the frequency dependent attenuation factor in dB for given air temperature, humidity, and pressure conditions. At distance p, the magnitudes of $X(f)$ should be attenuated by $D(f,p)$ decibels:

$$D(f,p) = p \cdot a(f) \quad (5)$$

where $a(f)$ is the frequency-dependent attenuation, which will have an impact on the brightness of the sound (higher frequencies being more attenuated than lower ones).
More precisely, the total absorption in decibels per meter $a(f)$ is given by a rather complicated formula:

$$\frac{a(f)}{P} \approx 8.68 \cdot F^2 \left\{ 1.84 \cdot 10^{-11} \left(\frac{T}{T_0}\right)^{1/2} P_0 + \left(\frac{T}{T_0}\right)^{-5/2} \right.$$
$$\left[0.01275 \cdot e^{-2239.1/T} / \left[F_{r,O} + \left(F^2 / F_{r,O} \right) \right] \right. \quad (6)$$
$$\left. \left. +0.1068 \cdot e^{-3352/T} / \left[F_{r,N} + \left(F^2 / F_{r,N} \right) \right] \right] \right\}$$

where $F = f / P$, $F_{r,O} = f_{r,O} / P$, $F_{r,N} = f_{r,N} / P$ are frequencies scaled by the atmospheric pressure P, and P_0 is the reference atmospheric pressure (1 atm), f is the frequency in Hz, T is the atmospheric temperature in Kelvin (K), T_0 is the reference atmospheric temperature (293.15K), $f_{r,O}$ is the relaxation frequency of molecular oxygen, and $f_{r,N}$ is the relaxation frequency of molecular nitrogen. See [10] for details.

B. Binaural Spatialization

In binaural listening conditions using headphones, the sound from each earphone speaker is heard only by one ear. Thus the encoded spatial cues are not affected by any cross-talk signals between earphone speakers.
To spatialize a sound source to an expected azimuth θ, for each short-term spectrum X, we compute the pair of left X_L and right X_R spectra from the spatial cues corresponding to θ, using Equations (1) and (3), and:

$$\begin{cases} X_L(t,f) = X(t,f) \cdot 10^{+\Delta_a(f)/2} e^{+j\Delta_\Phi(f)/2} \\ X_R(t,f) = X(t,f) \cdot 10^{-\Delta_a(f)/2} e^{-j\Delta_\Phi(f)/2} \end{cases} \quad \text{(7) and (8)}$$

(because of the symmetry among the left and right ears), where Δ_a and Δ_Φ are given by:

$$\begin{cases} \Delta_a(f) = ILD(\theta, f)/20 \\ \Delta_\Phi(f) = ITD(\theta, f) \cdot 2\pi f \end{cases} \quad \text{(9) and (10)}$$

The control of both amplitude and phase should provide better audio quality than amplitude-only spatialization1 (see below).
Indeed, we reach a remarkable spatialization realism through informal listening tests with AKG K240 Studio headphones. The main problem which remains is the classic front / back confusion.

C. Multi-Loudspeaker Spatialization

In a stereophonic display, the sound from each loudspeaker is heard by both ears. Thus, the stereo sound is filtered by a matrix of four transfer functions $\left(C_{i,j}(f,\theta)\right)$ between loudspeakers and ears (see Figure 4). Here, we generate the paths artificially using the binaural model. The best panning coefficients under CIPIC conditions for the pair of speakers to match the binaural signals at the ears (see Equations (7) and (8)) are then given by:

$$\begin{aligned} K_L(t,f) &= C \cdot \left(C_{RR} H_L - C_{LR} H_R\right) \\ K_R(t,f) &= C \cdot \left(-C_{RL} H_L + C_{LL} H_R\right) \end{aligned} \quad \text{(10) and (11)}$$

with the determinant computed as:

$$C = 1/\left(C_{LL} C_{RR} - C_{RL} C_{LR}\right) \quad \text{(12)}$$

In extreme cases where $|C| = 0$ (or close to zero) at any frequency, the matrix C is ill-conditioned, and the solution becomes unstable. To avoid unstable cases, attention should be paid during the loudspeakers configuration stage, before live diffusion.

During diffusion, the left and right signals (Y_L, Y_R) to feed left and right speakers are obtained by multiplying the short-term spectra X with K_L and K_R, respectively:

$$Y_L(t,f) = K_L(t,f) \cdot X(t,f)$$
$$Y_R(t,f) = K_R(t,f) \cdot X(t,f) \quad \text{(13) and (14)}$$

In a setup with many speakers we use the classic pairwise paradigm, consisting in choosing for a given source only the two speakers closest to it (in azimuth): one at the left of the source, the other at its right.

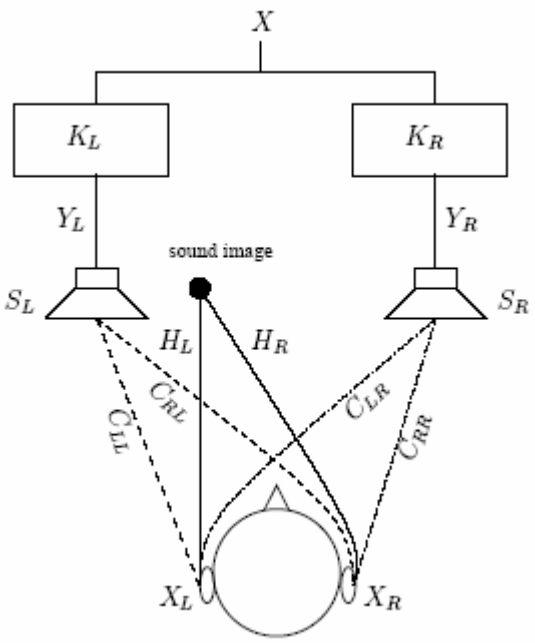


Fig. 4 Stereophonic loudspeaker display

D. Analysis of Panning Coefficients

We used the speaker pair (−30◦, +30◦) to compute the panning coefficients at any position (between the speakers) with the two techniques: our approach and the classic vector-based amplitude panning (VBAP) approach [3]. VBAP was elaborated under the assumption that the incoming sound is different only in amplitude, which holds for frequencies up to 600Hz. In fact, by controlling correctly the amplitudes of the two channels, it is possible to produce resultant phase and amplitude differences for continuous sounds that are very close to those experienced with natural sources. We restrict our comparisons to the [0, 800]Hz frequency band.

1) Comparisons of Panning Coefficients: The panning coefficients of the two approaches are very similar until 600Hz (see Figure 5), and can differ significantly above. In fact, our coefficients are complex values, and their imaginary parts can contribute in a significant way (see Figure 6).

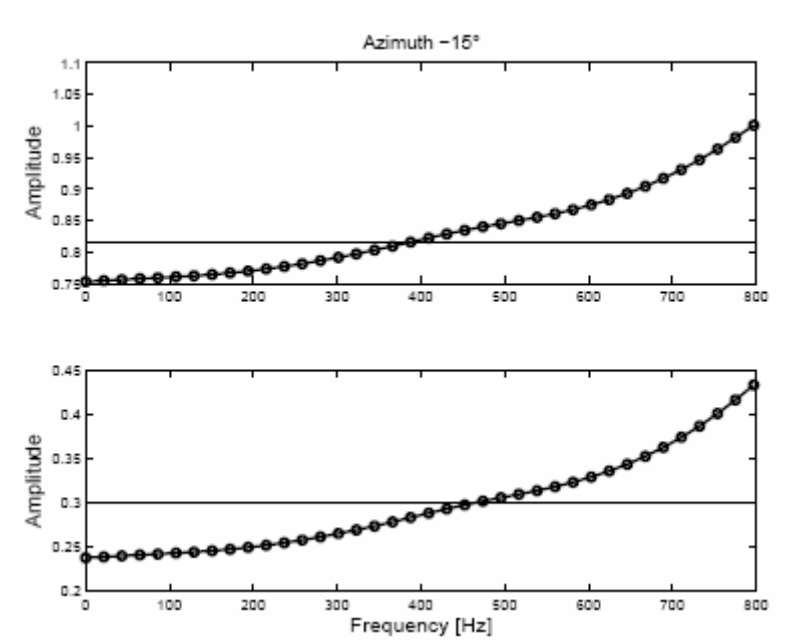


Fig. 5 Amplitude of the panning coefficients for the left (top) and right (bottom) channels

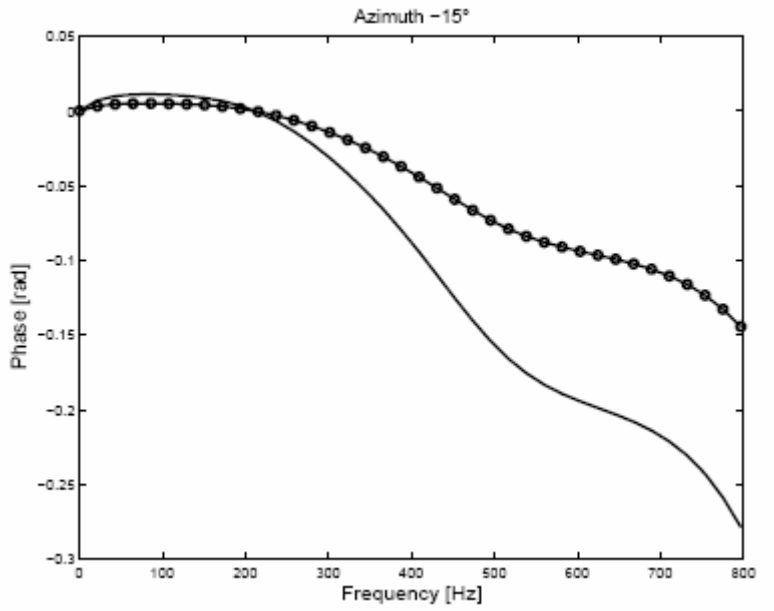


Fig. 6 Phase of the panning coeffisients

2) Comparisons of the Ratio of Panning Coefficients: Generally, inter-channel differences are perceptually more relevant (*e.g.* ILD, ITD) than absolute values. Given the left and right panning coefficients, K_L and K_R, we compute the *panning level difference* (PLD):

$$PLD = 20\log_{10}\left|\frac{K_L}{K_R}\right| \quad (15)$$

We computed the absolute difference between the PLDs of both VBAP and our approach. The maximal PLD difference (in the considered frequency band) has a linear trend, and its maximum does not exceed 3dB. Thus, the two approaches seem to be consistent in the [0, 800]Hz band (see Figure 7). For higher frequencies, the new approach should yield better results, as confirmed perceptively in our preliminary and informal listening tests.

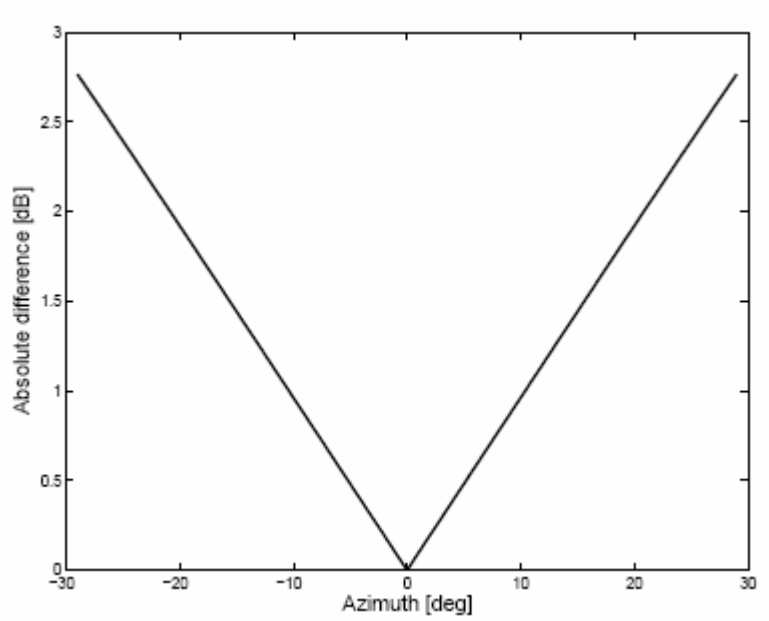


Fig. 7 Maximum difference per azimuth

2.2 Localization Method

A. Azimuth Estimation

In Auditory Scene Analysis (ASA), ILDs and ITDs are the most important cues for source localization. Lord Rayleigh mentioned in his Duplex Theory that the ILDs are more prominent at high frequencies (where phase ambiguities are likely to occur) whereas the ITDs are crucial at low frequencies (which are less attenuated during their propagation).
Obtaining an estimation of the azimuth based on the ILD information (see Equation (2)) is just a matter of inverting Equation (1):

$$\theta_L(t,f) = \arcsin\left(\frac{ILD(t,f)}{\alpha(f)}\right) \quad (16)$$

Similarly, using the ITD information (see Equation (4)), to obtain an estimation of the azimuth candidate for each p, we invert Equation (3):

$$\theta_{T,p}(t,f) = \arcsin\left(\frac{c \cdot ITD_p(t,f)}{r \cdot \beta(f)}\right) \quad (17)$$

The $\theta_L(t,f)$ estimates are more dispersed, but not ambiguous at any frequency, so they are exploited to find the right modulo coefficient p that unwraps the phase. Then the $\theta_{T,p}(t,f)$ that is nearest to $\theta_L(t,f)$ is validated as the final θ estimation for the considered frequency bin, since it exhibits a smaller deviation:

$$\theta(t,f) = \theta_{T,m}(t,f) \quad (18)$$

With $m = \arg\min_p \left|\theta_L(t,f) - \theta_{T,p}(t,f)\right|$ (19).

Practically, the choice of p can be limited among two values $\left(\lfloor p_r \rfloor, \lceil p_r \rceil\right)$, where

$$p_r = \left(f \cdot ITD(\theta_L, f) - \frac{1}{2\pi} \angle \frac{X_L(t,f)}{X_R(t,f)} \right) \quad (20)$$

An estimate of the azimuth of the source can be obtained as the peak in an energy-weighted histogram (see [9]). More precisely, for each frequency bin of each discrete spectrum, an azimuth is estimated and the power corresponding to this bin is accumulated in the histogram at this azimuth. For the corresponding bin frequency f, the power $\left|X(f)\right|^2$ is estimated by inverting Equations (7) and (8) for the left and right spectra, respectively, then the square of the estimate of the loudest – supposedly most reliable – channel is retained for the power estimate.

Thus, we obtain a power histogram as shown in Figure 8. This histogram is the result of the localization of a Gaussian white noise of 0.5s spatialized at azimuth −45◦. On this figure, we can clearly see two important local maxima (peaks), one around azimuth −45◦, the other at azimuth −90◦. The first (and largest) one corresponds to the sound source; the second one is a spurious peak resulting from extreme ILDs (a problem we have to solve in our future research).

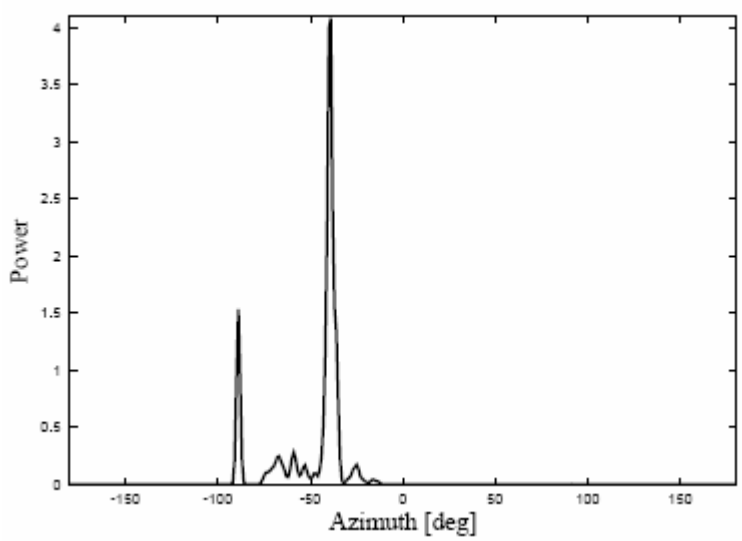


Fig. 8 Histogram obtained with a source at azimuth -45º

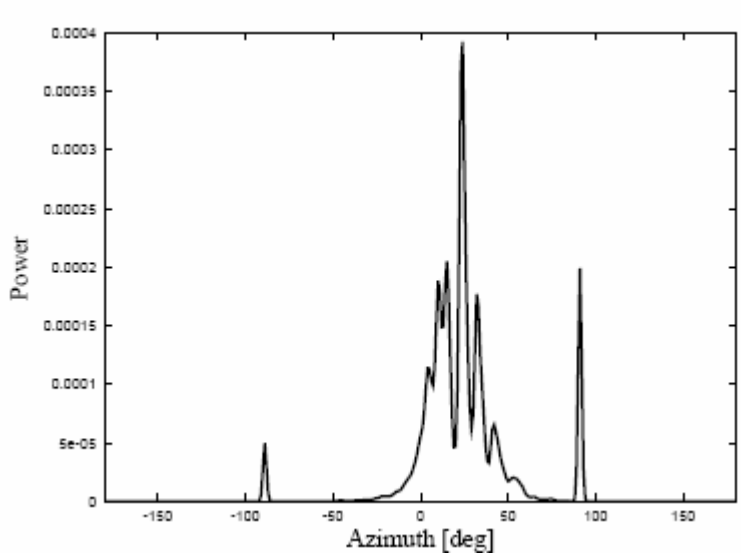


Fig. 9 Histogram obtained with a real source positioned at azimuth 30º

For our localization tests, we spatialized a Gaussian white noise using convolutions with the HRTFs of the KEMAR manikin (see [2]), since they were not part of the database used for the learning of our model coefficients and thus should give results closer to those expected with a real – human – listener. Indeed, in our first experiments with real listeners (see Figure 9), the same trends as in Figure 8 were observed: a rather broader histogram but still with a local maximum close to the azimuth of the sound source, plus spurious maxima at extreme azimuths ±90◦.

To verify the precision of the estimation of the azimuth, we spatialized several noise sources at different azimuths in the horizontal plane, between −80◦ and +80◦, and we localized them using the proposed method. The results are shown in Figure 10. We observe that the absolute azimuth error is less than 5◦ in the [−65,+65]◦ range.

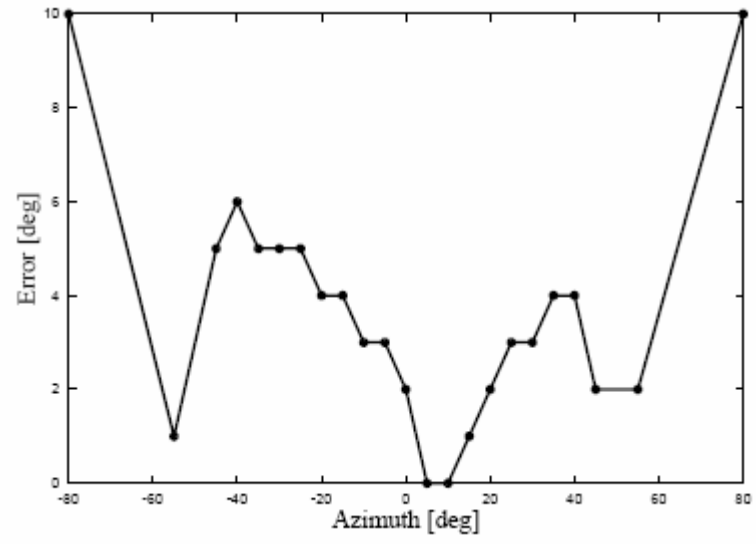


Fig. 10 Absolute error of the localization of the azimuth

In real reverberant environments, due to more superpositions at the microphones, an amplitude-based method is not really adapted; in contrast, generalized crosscorrelation based ITD estimation should be more robust.

B. Distance Estimation

As a reference signal for distance estimation, we use a Gaussian white noise spatialized at azimuth zero, since pure tones are not suitable for distance judgments. The distance estimation relies on the quantification of the spectral changes during the sound propagation in the air.
To estimate the amplitude spectrum, we first estimate the power spectral density of the noise using the Welch's method. More precisely, we compute the mean power of the short-term spectra over L frames, then take its square root, thus:

$$|X| = \sqrt{\frac{1}{L}\sum_{l=-(L-1/2)}^{l=+(L-1/2)} |X_l|^2} \quad (21)$$

In our experiments, we consider $L = 21$ frames of $N = 2048$ samples, with an overlap factor of 50% (and with a CD-quality sampling rate of 44.1kHz, thus the corresponding sound segment has a length < 0.5s).
Then we use this amplitude spectrum to compute the spectral centroid:

$$C = \frac{\sum_f f \cdot |X(f)|}{\sum_f |X(f)|} \quad (22)$$

The spectral centroid moves towards low frequencies when the source moves away from the observer. The related perceptive brightness is an important distance cue.
We know the reference distance since the CIPIC speakers were positioned on a 1-m radius hoop around the listener. By inverting the logarithm of the function of Figure 11, obtained thanks to the ISO 9613-1 norm and Equations (5), (6), and (22), we can propose a function to estimate the distance from a given spectral centroid:

$$p(\log(C)) = -38.89044C^3 + 1070.33889C^2 - 9898.69339C + 30766.67908 \quad (23)$$

given for the air at 20◦ Celsius temperature, 50% relative humidity, and 1 atm pressure.

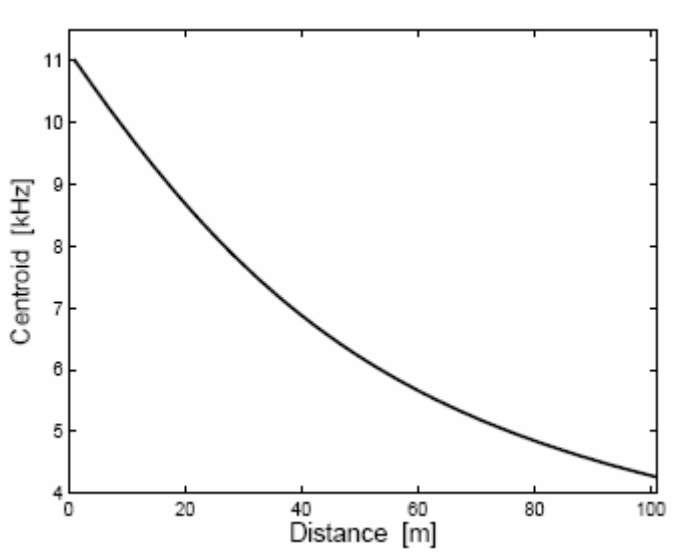


Fig. 11 Spectral centroid (related to perceptive brightness)

Up to 25m, the maximum distance error is theoretically less than 4mm, if the noise power spectral density is known. However, if the amplitude spectrum has to be estimated using Equation (21), then the error is greater, though very reasonable until 50m. Figure 12 shows the results of our simulations for Gaussian white noise spatialized at different distances in the [0, 100]m range.

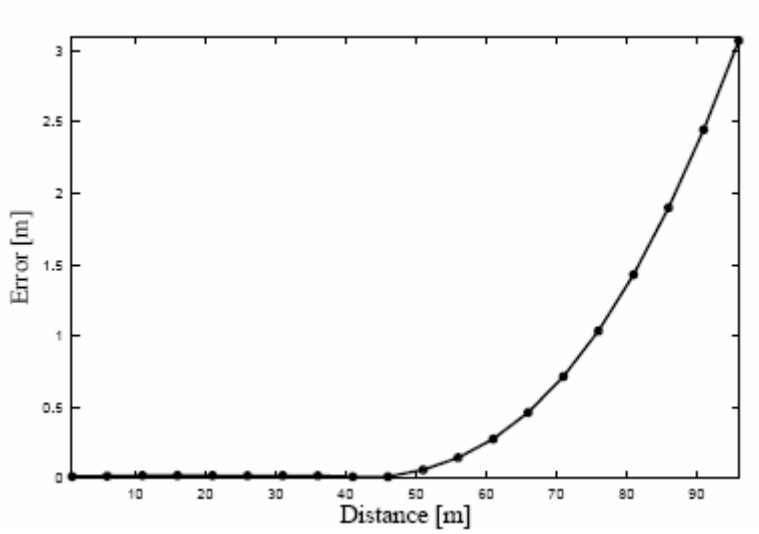


Fig. 12 Absolute error of the localization of the distance

3 SOURCE CODE

The RetroSpat system is being implemented as a realtime musical software under the GNU General Public License (GPL). The actual implementation is based on C++, Qt42, JACK3, FFTW4 and works on Linux and MacOS X.
Currently, RetroSpat implements the described methods (*i.e.* localization and spatialization) in two different modules: *RetroSpat Localizer* for speaker setup detection and *RetroSpat Spatializer* for the spatialization process. We hope to merge the two functionalities in one unique software soon.

Currently, RetroSpat implements the described methods in two different modules: *RetroSpat Localizer* for speaker setup detection and *RetroSpat Spatializer* for the spatialization process.

A. RetroSpat Localizer

RetroSpat Localizer (see Figure 13) is in charge of the automatic detection of the speakers configuration. It also allows the user to interactively edit a configuration, which has been just detected or loaded from an XML file.
The automatic detection of the positions (azimuth and distance) of the speakers connected to the soundcard is of great importance to adapt to new speaker setups. Indeed, it will be one of the first actions of the interpreter in a new environment.

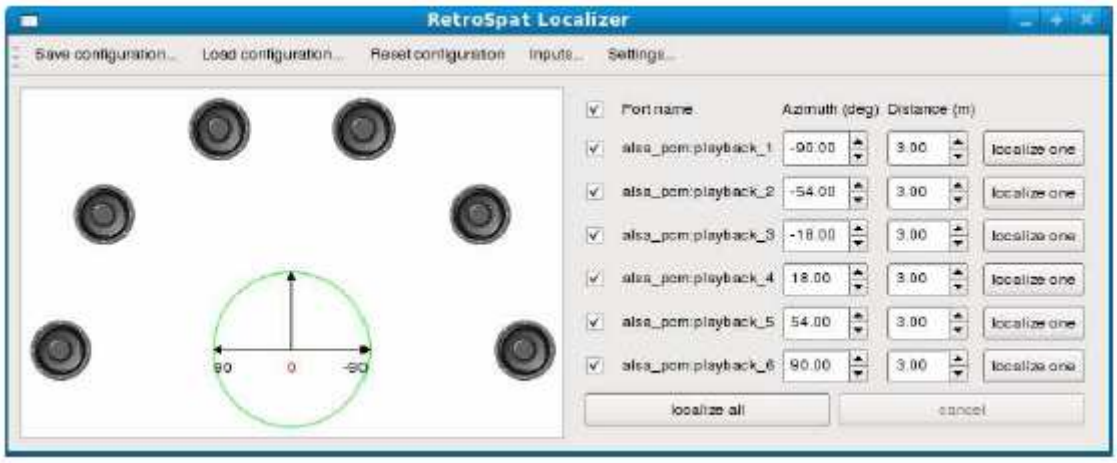


Fig. 13 Retrospat Localizer graphical user interface with a 6-speaker configuration

For room calibration, the interpreter carries headphones with miniature microphones encased in earpieces (see Figure 15, where Sennheiser KE4-211-2 microphones have been inserted in standard headphones). The interpreter orients the head towards the desired zero azimuth. Then, each speaker plays in turn a Gaussian white noise sampled at 44.1kHz. The binaural signals recorded from the ears of the musician are transferred to the computer running RetroSpat Localizer. Each speaker is then localized in azimuth and distance. The suggested configuration can be adjusted or modified by the interpreter according to the rooms characteristics.

Fig. 14 The “phonocasque” used for the binaural recordings

B. RetroSpat Spatializer

For sound spatialization, mono sources are loaded in RetroSpat, parameterized, and then diffused. The settings include the volume of each source, the initial localization, the choice of special trajectories such as circle, arc, etc. A loudspeaker-array configuration is the basic element for the spatialization (see Section VI-A).
The snapshot of Figure 14 depicts a 7-source mix of instruments and voices (note icons), in a 6-speaker frontfacing configuration (loudspeaker icons), obtained from RetroSpat Localizer.
During the diffusion, the musician can interact individually with each source of the piece, change its parameters (azimuth and distance), or even remove / insert a source from / into the scene. In this early version, the interaction with RetroSpat is provided by a mouse controller.
Thanks to an efficient implementation using the JACK sound server, RetroSpat can diffuse properly simultaneous sources even within the same speaker pair (see Figure 14, three sources in speaker pair (2,3)). All the speaker pairs have to stay in synchrony. To avoid sound perturbation, the Qt-based user interface runs in a separated thread with less priority than the core signal processing process. We tested RetroSpat on a MacBook Pro, connected to 8 speakers, through a MOTU 828 MKLL soundcard, and were able to play several sources without problems. However, further testing is needed to assess scalability limits.

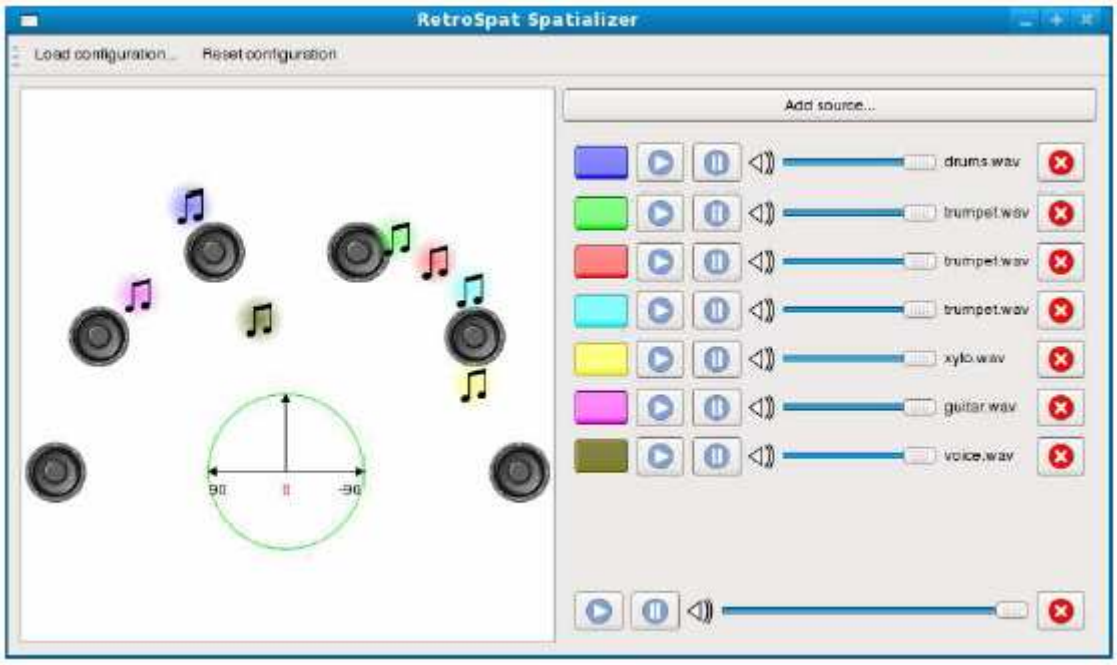


Fig. 15 Retrospat Spatializer graphical user interface, with 7 sources spatialized on the speaker setup presented on Figure 13

C. Musical Applications

In a live concert, the acousmatic musician interacts with the scene through a special (un)mixing console.
With RetroSpat, the musician has more free parameters on one single controller (mouse):

- only mouse movement to control simultaneously the azimuthal and distance location;
- mono sources as inputs;
- many sources can be spatialized to different locationsat the same time;
- a dynamic visualization of the whole scene (sourceapparition, movement, speed, etc.) is provided.

We believe that RetroSpat should greatly simplify the interpreter interactions and thus should allow him / her to focus more on the artistic performance.

3.1 Programme Language used

The programme language used to implement the *Localization* and *Spazialization* Methods is MATLAB (version 7.1.0.246(R14) Service Pack 3).

MATLAB is a numerical computing environment and fourth generation programming language. Developed by The MathWorks, MATLAB allows matrix manipulation, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs in other languages. Although it is numeric only, an optional toolbox uses the MuPAD symbolic engine, allowing access to computer algebra capabilities. An additional package, Simulink, adds graphical multidomain simulation and Model-Based Design for dynamic and embedded systems.

3.2 Code Structure

In this subsection it is tried how the code is structured as well as what functions are included in each folder. Next figure (Fig. 16) shows the code structure:

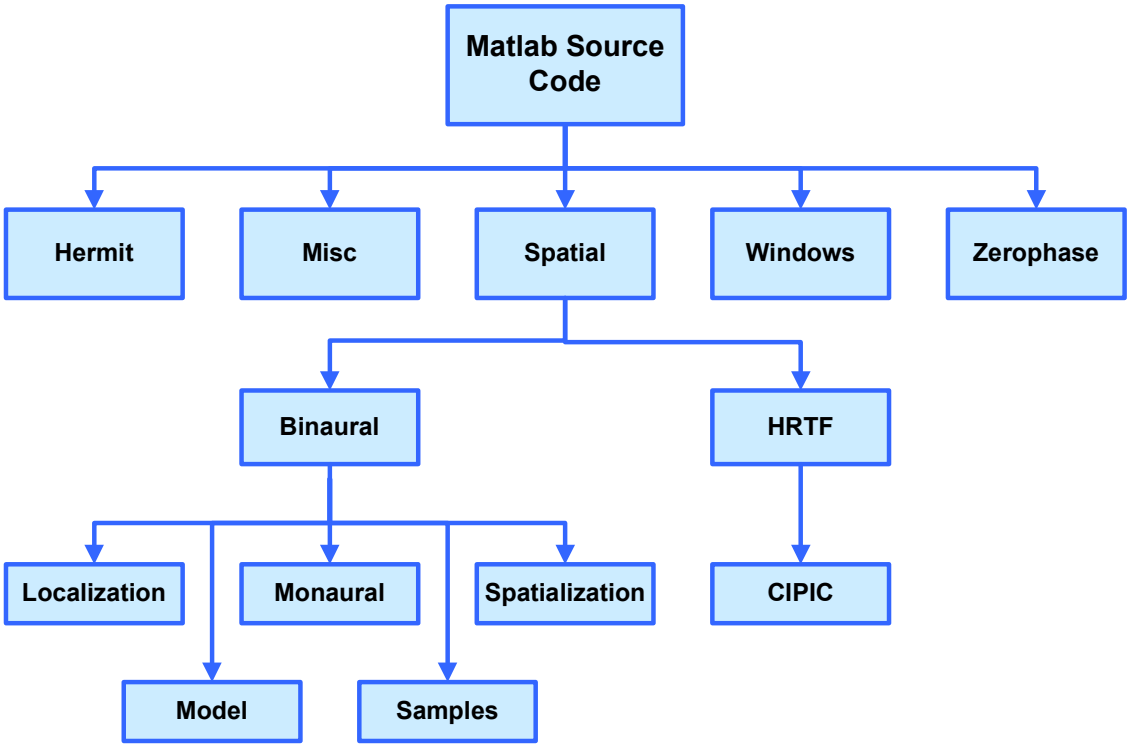


Fig. 16 Code Structure

The code is composed of different areas corresponding to the aim of the functions stored in each folder. So the first folder named *Hermit* corresponds to the hermitic matrix functions used in signal processing when dealing with Fourier transforms.

The *Misc* area contains all the functions corresponding to mathematics operations as the fft (fast Fourier transform), conversion functions as transform radians to degrees or readjusting functions as fixing a new time axis.

The *Spatial* folder corresponds with the suitable functions of the RetroSpat System. It is divided in two sections. One referred to the Head-Related Transfer Functions (HRTF) using the CIPIC HRTF Database for signal processing to audio and acoustics. And the other container of all the functions related to the model (binaural and monaural), to the localization and spatialization method and to the storage of the samples.

The *Windows* section includes all the functions referred to the Hann window function used to make the spectral analysis of the equations.

At last, there is the *Zerophase* folder used to get the zero-phase response and the frequency vector associated at this response.

For seeing the relation between function see appendix 1.

3.3 Code Architecture

Next, there is a diagram (see Fig. 17) which tries to summarise the aim of the source code. It can be seen that the intermediate step (when the spectrum and azimuth are localized) is the point where the RetroSpat is focused because with these parameters it can be guess the positions of the sound source and consequently it can be controlled and moved.

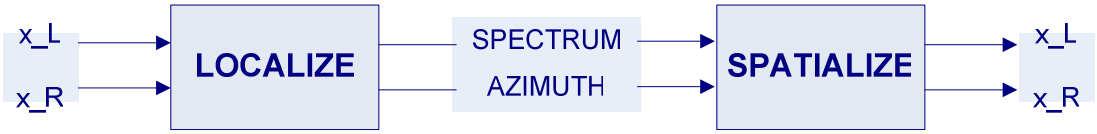


Fig. 17 Code Diagram

Let's do a more accurate diagram in order to understand which the steps in each method are. First, there is a chain of the localize method (Fig.18) and the Spatialize (Fig.19).

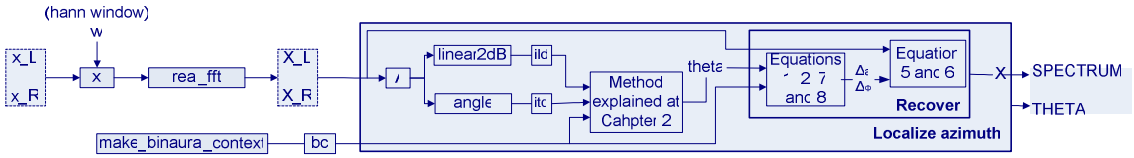


Fig. 18 Localize Method

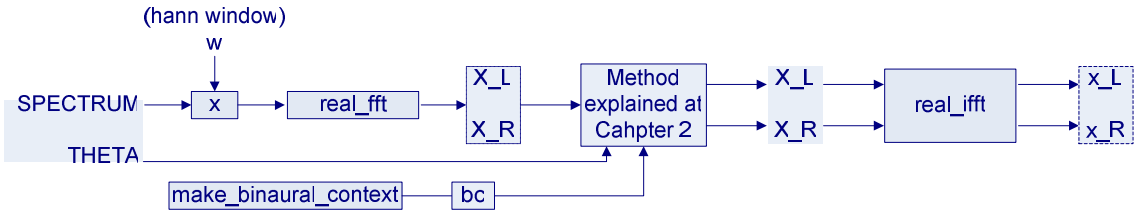


Fig. 19 Spatialize Method

3.3.1 Spatialization Method Architecture

In order to make the code comprehension easier in this section it has tried to summarized the whole process of the *test_spatialize* in a flowchart input/output (Fig. 20). The number 16 corresponds to the number of bits per sample.

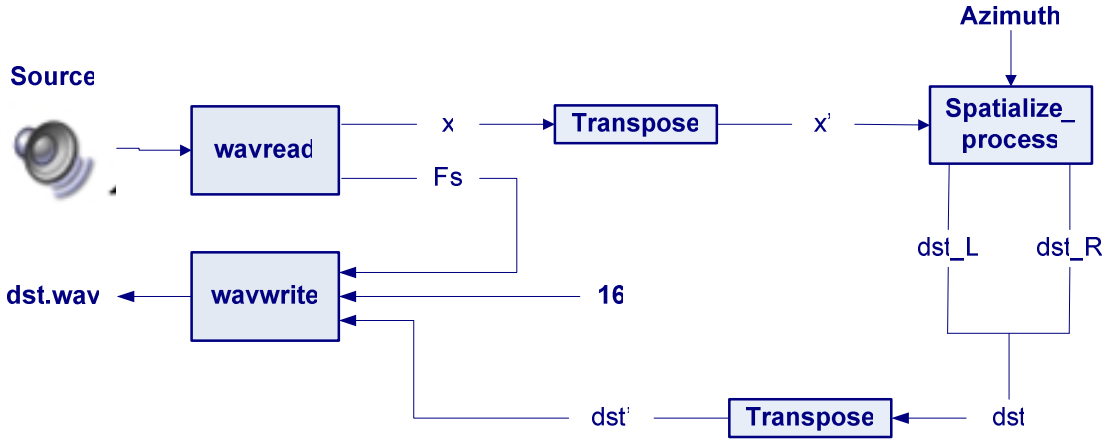


Fig. 20 Spatialize Method Architecture

3.3.2 Localization Method Architecture

In this section it has done the same as the section before: try to synthesize as much as possible the functions used in the source code. For the *Localization Method* it has been necessary three different schemas because of the complexity and extension of the functions.
First, it is done the diagram of the main function in the method: *test_localize* (Fig. 21).

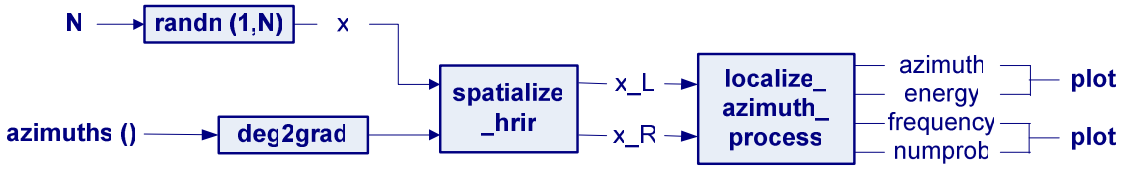


Fig. 21 Test_localize

As it can be seen there are two functions that have to be described next. The first one is *spatialize_hrir* where with the sound source and the azimuths used are obtained the two channels (left and right) of the sound source separately (Fig. 22).

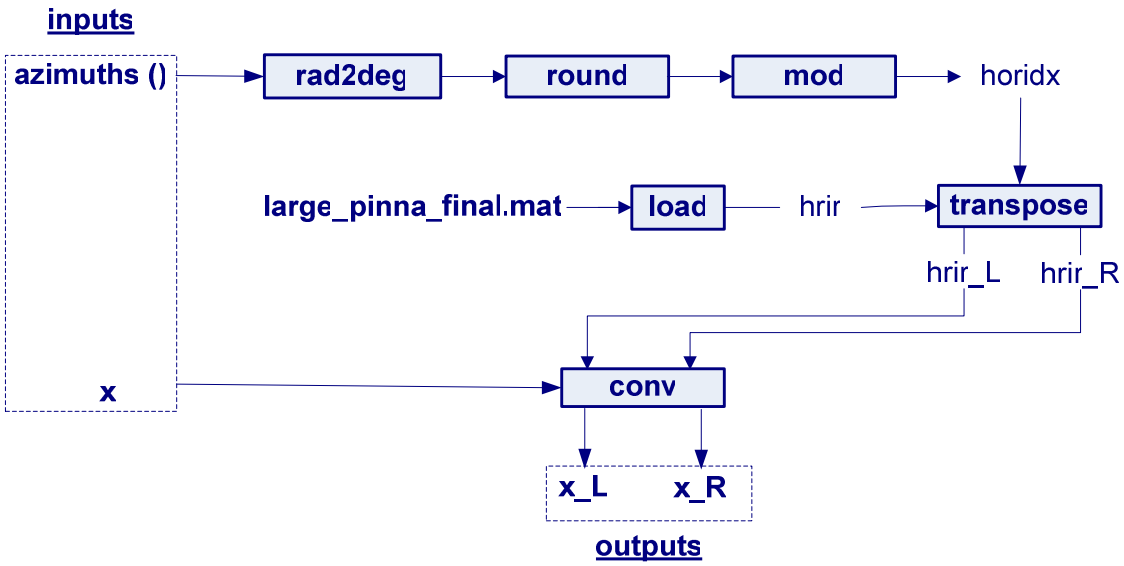


Fig.22 Spatialize_hrir

For the schema of the *localization_azimuth_process* (see Fig. 23) is necessary to explain why the power is calculated with $\frac{2|X|^2}{N\sum w^2}$. All that is known is that the signal x is multiplied by the w of the hann window, so:

$$\sum_N (x \times w)^2 = \frac{1}{N}\sum_N |X|^2$$ Parseval's theorem

$$\sum_N (x \times w)^2 = \sum_N x^2 \times \sum_N w^2$$ The signal is 0 mean

$$\frac{1}{N}\sum_N |X|^2 = \frac{2}{N}\sum_{N/2} |X|^2$$ It is real FFT

The box corresponding to the *localize_azimuth* was explained in section 3.3 (see Fig. 3).

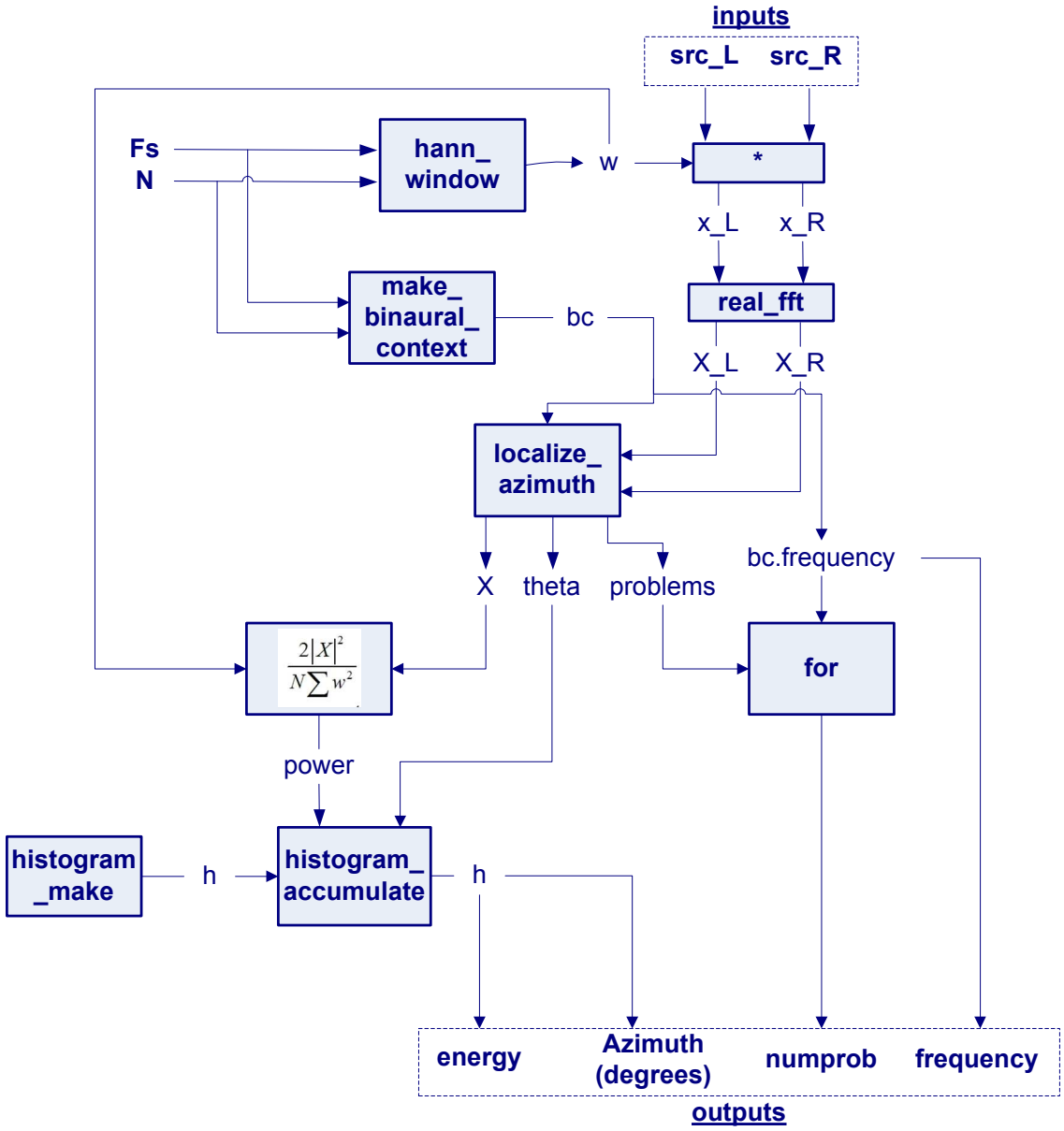


Fig. 23 Localization_azimuth_process

4 CURRENT FAULTS ON THE SOURCE CODE

The current code presents several faults but only in the *localization* method. The most part of them are warnings apparently easy to solve and the only error visible is a spurious peak around the azimuth -90º when plotting the power spectral of the sound source.

According to the warnings, the code had problems while it was calculated the azimuth θ using the functions *azimuth_from_ild* and *azimuth_from_itd,* and while calculating the *itd factor* with the *itd_from_phase* function.

The problem can be summarized in the calculation of those ratios:

$$ratio_{ild} = \frac{ILD(t,f)}{\alpha(f)}$$

$$ratio_{itd} = \frac{c \cdot ITD(t,f)}{r \cdot \beta(f)}$$

$$itd = \frac{phs + 2\pi p}{2\pi f}$$

It can be seen that the problem is related to the frequency vector or to the frequency-depended functions as $\alpha(f)$ and $\beta(f)$. So the frequency vector was checked in order to find out the problem and the problem was that the first position of the vector was zero. Then the corresponding modification was made and the problem with the warnings was solved.

However the error is not as easy to solve as the warnings. First, it is showed a figure that represents the histogram as a result of the localization of a Gaussian white noise of 0.5s spatialized at azimuth -45º as well the peak-error.

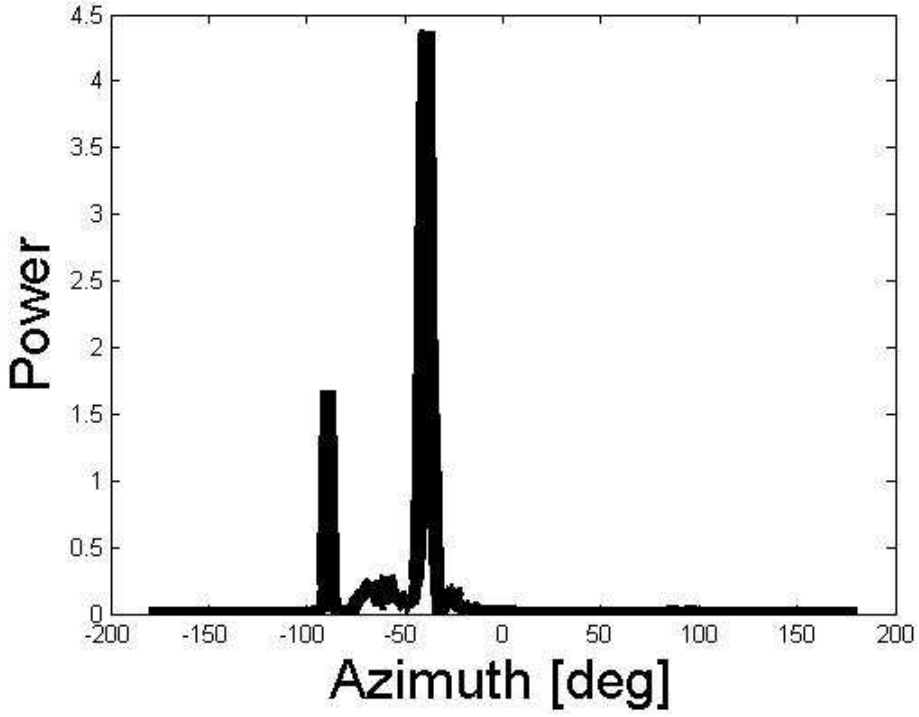


Fig. 24 Histogram at azimuth 45º

In the figure, it can clearly be seen two important peaks, one around azimuth -45º, corresponding to the sound source, and the other at azimuth -90º corresponding to a spurious peak.

So, in order to know which frequencies are problematic it is obtained a function sort by how many times there are problems with each frequency. Taking the problem as $abs(\theta) = 90º$. It has to take into account that the number of problems must be calculated for each k-value (azimuth) and for each H-value (length of the window used to obtain the FFT (*Fast Fourier Transform*) of the sound source).

The number of k-values and H-values are 25 and 19 respectively. So the maximum number of errors for each frequency will be consequently 475. In order to have a reliable graphic to generalise for all frequencies, all the values will be divided by the maximum number of error.

Next figure (Fig. 25) shows which frequencies are more problematic than others. It is possible to say that for the very low frequency there will be always problems with the spurious peak and for the rest of the frequencies there is a fluctuation between 0% and 40% of peak appearance.

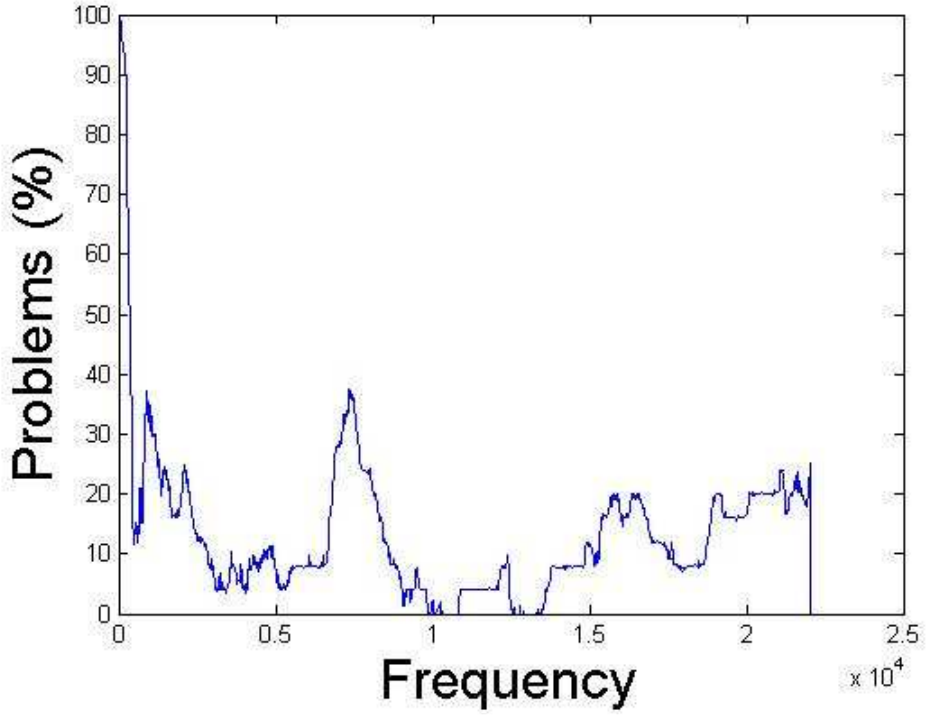


Fig. 25 Number of problems (%)

At first, it was believed that the problem was with the *ild factor* when it was combined to find an estimation of the azimuth θ because of the low frequency. The localization method uses the equations [9] and [10] to obtain $\theta_L(t,f)$ and $\theta_{T,p}(t,f)$ for each p-value [11].

The next figure shows the plot of the *ild factor* before it is combined with the *itd factor* to find the final azimuth θ estimation.

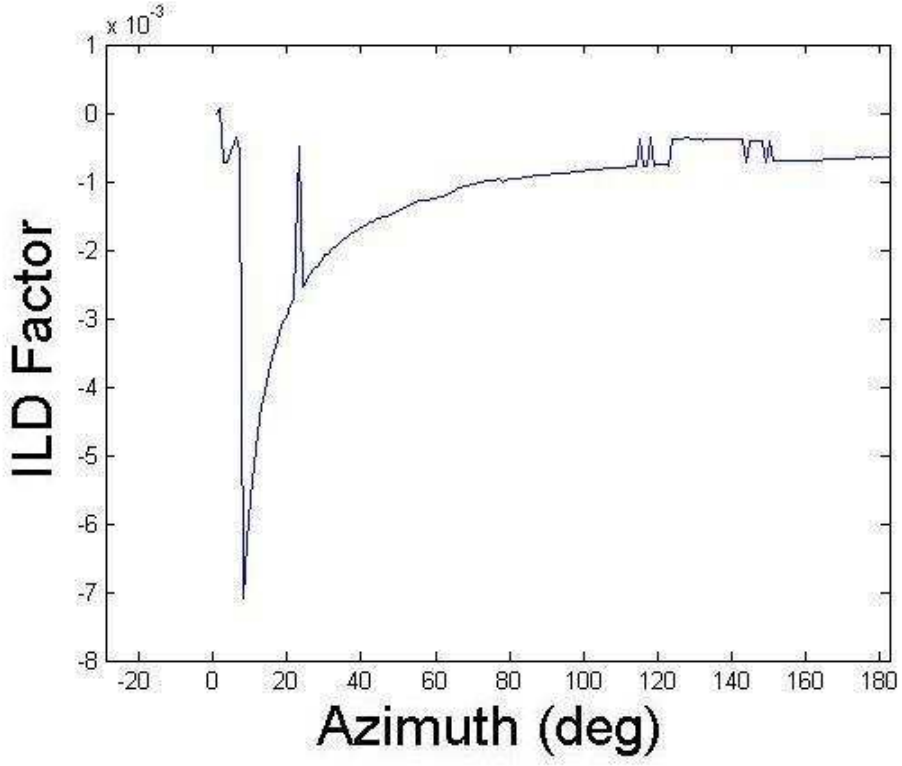


Fig. 26 ILD Factor

It can be seen that there is a problem with frequencies closer to zero. Then it is logical to think that avoiding the first samples and using only the *itd factor* to calculate the first samples of the azimuth θ it would be sufficient to correct the problem. So the changes made to code were only replacing the coefficient p (see Section 1) by zero for the 32 first samples.

Therefore, it is calculated again the number of errors and it can be seen that there is an improvement in the very low frequencies (see Figure 27).

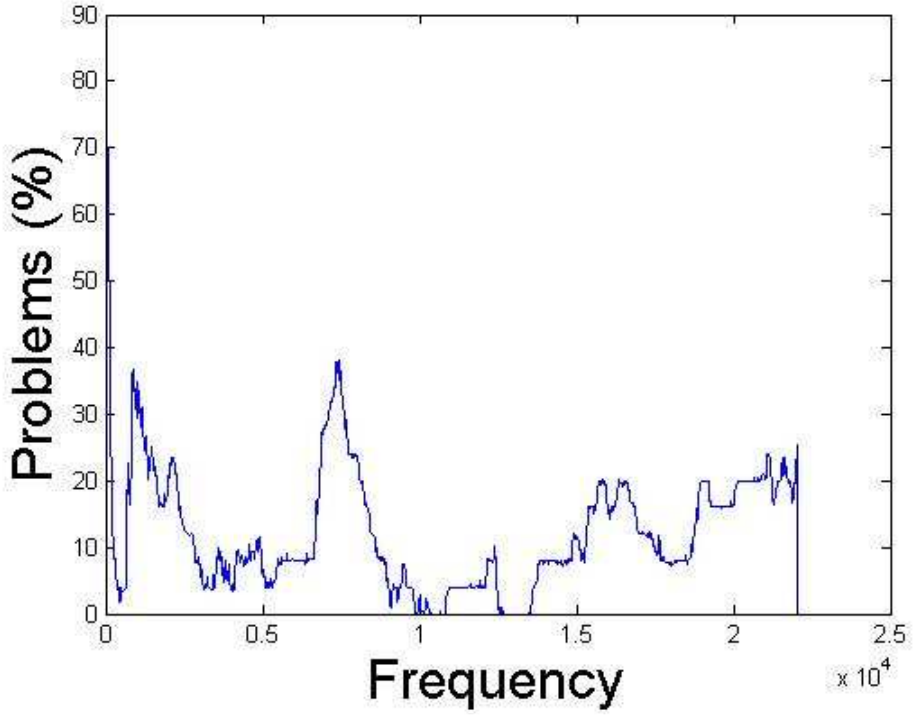


Fig. 27 Number of problems (%)

To be sure that there is an improvement a zoom of the two graphics (Fig. 25 and Fig. 27) is done. Here, it can be clearly seen that there is a decreasing in the number of error up to 85% and the frequencies with this probability of error have decreased too.

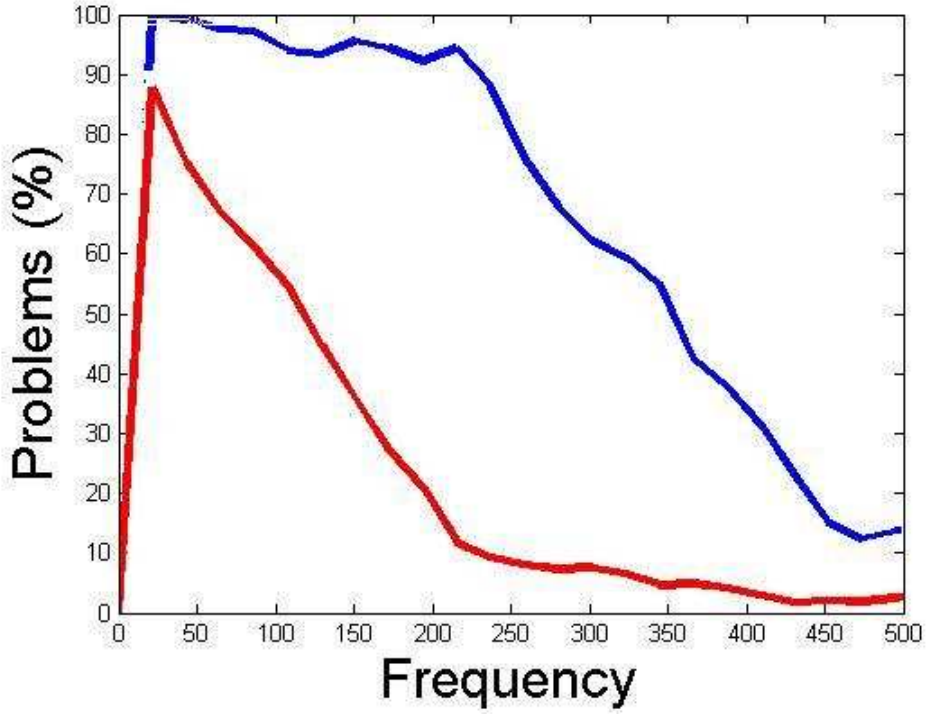


Fig. 28 Zoom of problems in the very low frequencies

Once it was ruled out that the problem was not with the *ild factor* it can be assured that the problem is with the $\alpha(f)$ factor. Maybe the problem is that $\alpha(f)$ is not large enough. To check it, the *line 20* (corresponding to the *ild factor*) of the function *make_binaural_context* is forced to a rescaling of 1.5.

Then the problems have reduced considerably (see Figure 29).

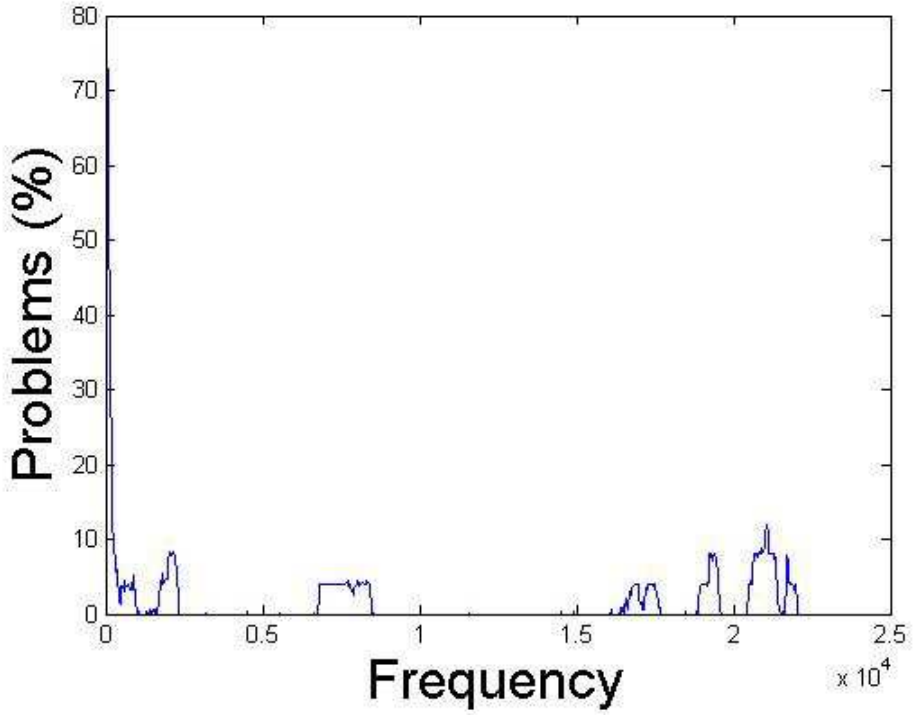


Fig. 29 Number of problems (%)

Despite this decreasing, there is a persisting problem with the very low frequency. But also, now there is another problem. If the equation [1] is taking into account it can be seen that $\frac{ILD}{\alpha(f)} = \sin(\theta)$, so if the $|\sin(\theta)| \leq 1$ then $\alpha(f)$ is not large enough to make the localization method true. The next figure (Fig. 30) shows the result in this case.

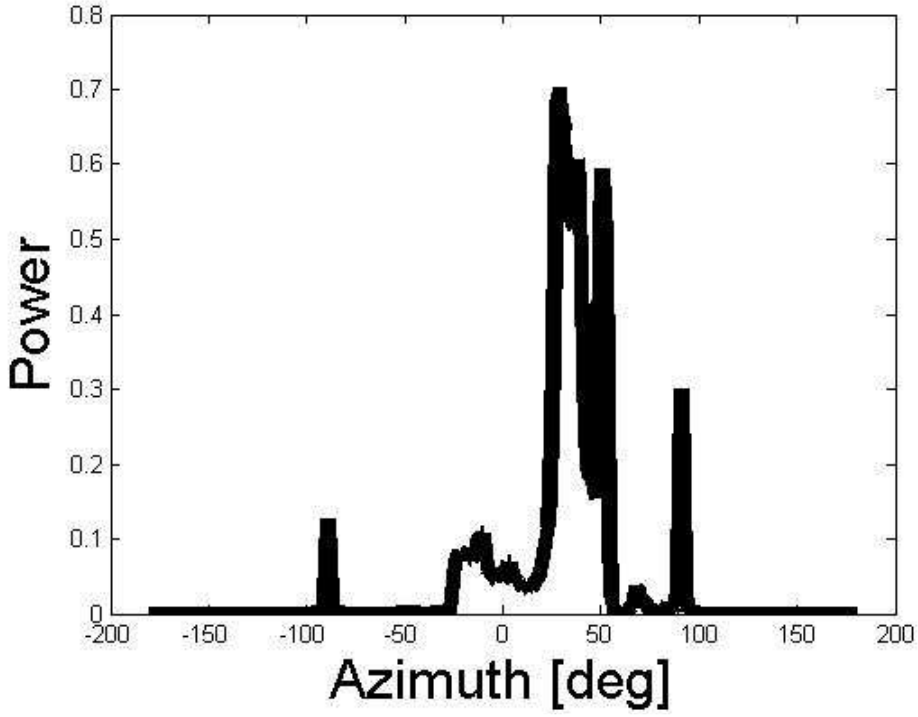


Fig. 30 Histogram at azimuth 45º

So it can be asserted that the spurious peak is a systematic error due to the statistical approximation made for the database.

5 Conclusions and future work

In this project, we have introduced a flexible multisource, multi-loudspeaker system: RetroSpat. This realtime system implements our proposed binaural to multiloudspeaker spatialization method. The system can also locate the loudspeakers azimuths and distances.
Several experiments at the SCRIME studio on an octophonic setup justify the utility of the system for live performance by composers of electroacoustic music. Next, we should enhance the source localization in real – reverberant – environments, and possibly evolve to source control through gesture or a more intuitive hardware controller. Also, a major scientific challenge would be to separate the different sources present in a binaural mix (for a semi-automatic diffusion from a compact disc as support).

REFERENCES

[1] H. Viste, “Binaural Localization and Separation Techniques,” Ph.D. dissertation, ´Ecole Polytechnique F´ed´erale de Lausanne, Switzerland, 2004.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF Database,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 2001, pp. 99–102.

[3] V. Pulkki, “Virtual Sound Source Positioning using Vector Base Amplitude Panning,” *Journal of the Acoustical Society of America*, vol. 45, no. 6, pp. 456–466, 1997.

[4] J. W. Strutt (Lord Rayleigh), “On the Acoustic Shadow of a Sphere,” *Philosophical Transactions of the Royal Society of London*, vol. 203A, pp. 87–97, 1904.

[5] ——, “Acoustical Observations I,” *Philosophical Magazine*, vol. 3, pp. 456–457, 1877.

[6] J. Blauert, *Spatial Hearing*, revised ed. Cambridge, Massachusetts: MIT Press, 1997, translation by J. S. Allen.

[7] R. S. Woodworth, *Experimental Psychology*. New York: Holt,
1954.

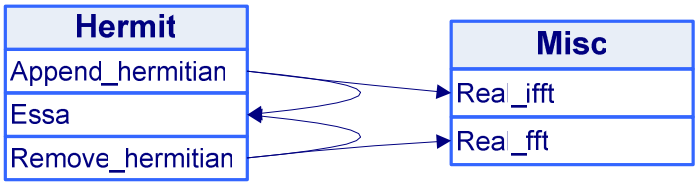
[8] G. F. Kuhn, “Model for the Interaural Time Differences in the Azimuthal Plane,” *Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977.

[9] J. Mouba and S. Marchand, “A Source Localization / Separation/ Respatialization System Based on Unsupervised Classification of Interaural Cues,” in *Proceedings of the Digital Audio Effects (DAFx) Conference*, Montreal, 2006, pp. 233–238.

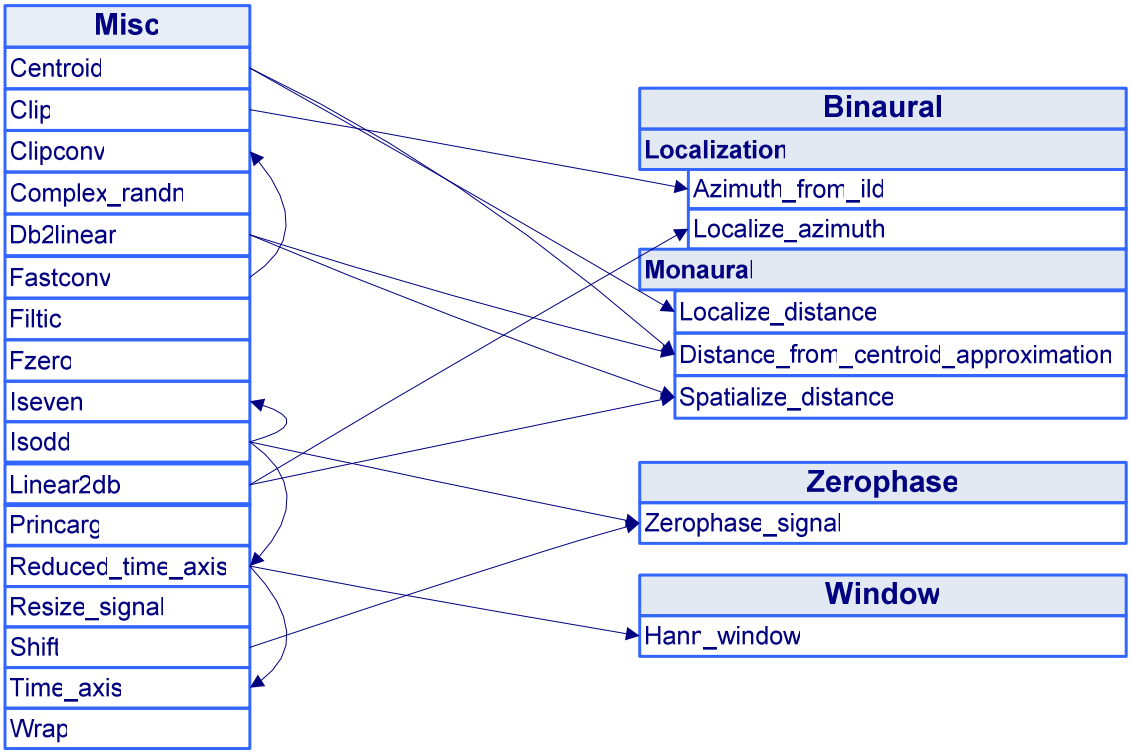
[10] H. Bass, L. Sutherland, A. Zuckerwar, D. Blackstock, and D. Hester, “Atmospheric Absorption of Sound: Further Developments,” *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 680–683, 1995.

APPENDIX

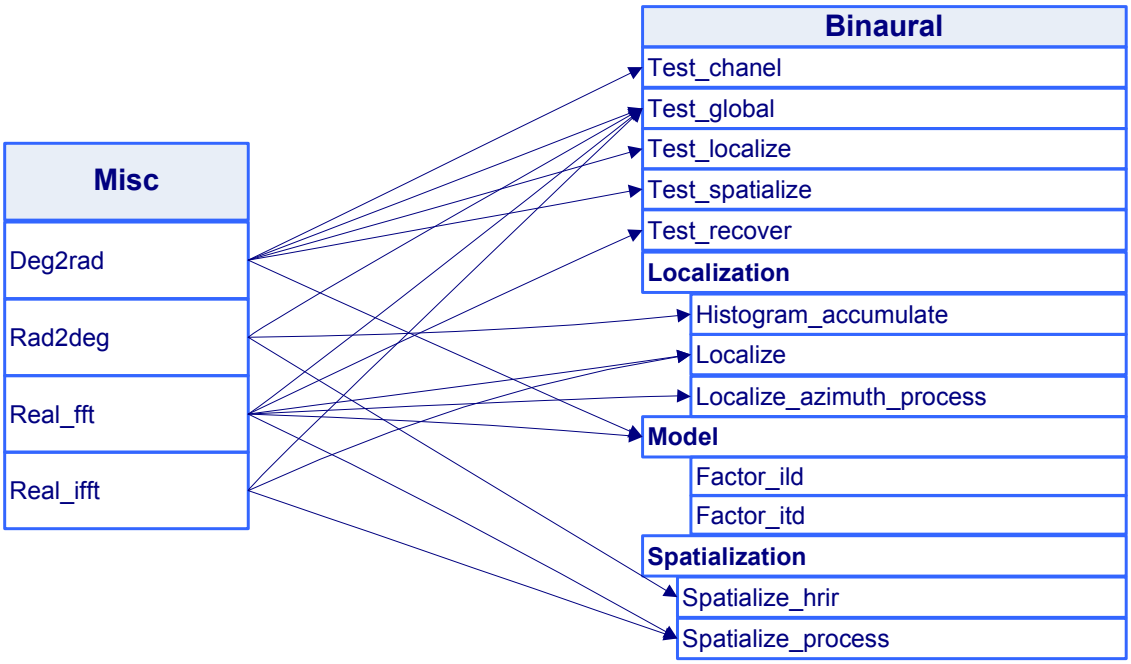
1. *Hermit* Folder



2. *Misc* Folder

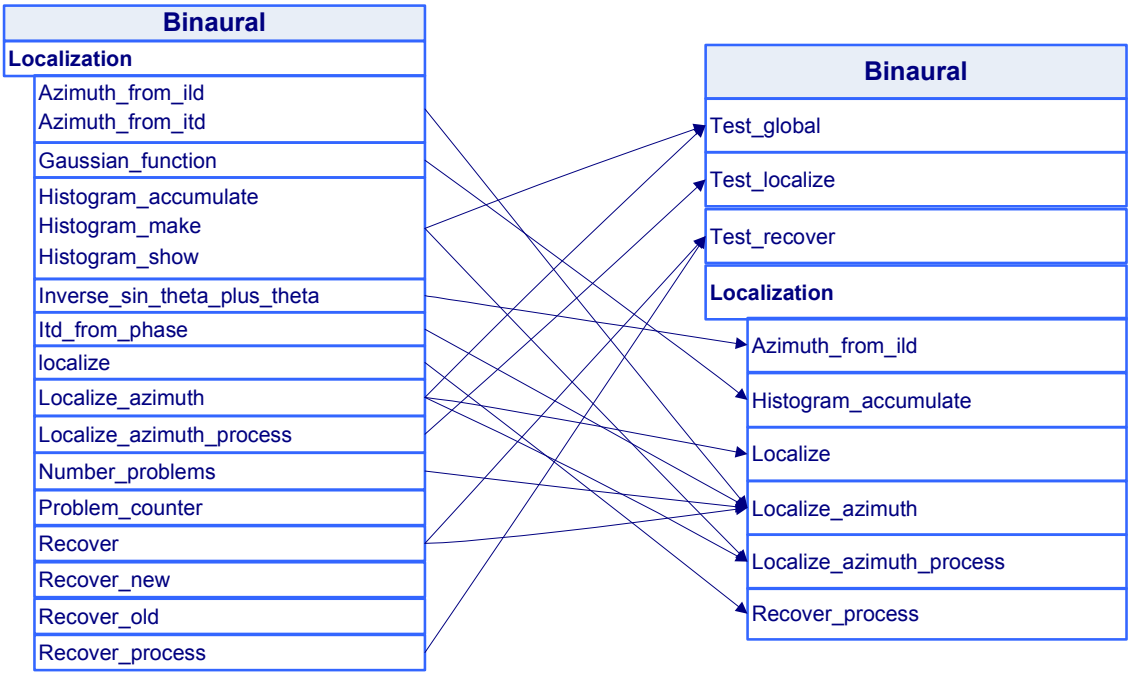


It is important to not making a mistake in the next figure it is necessary to notice that the narrows point at the *Model* cell only refers to the functions called afterwards.

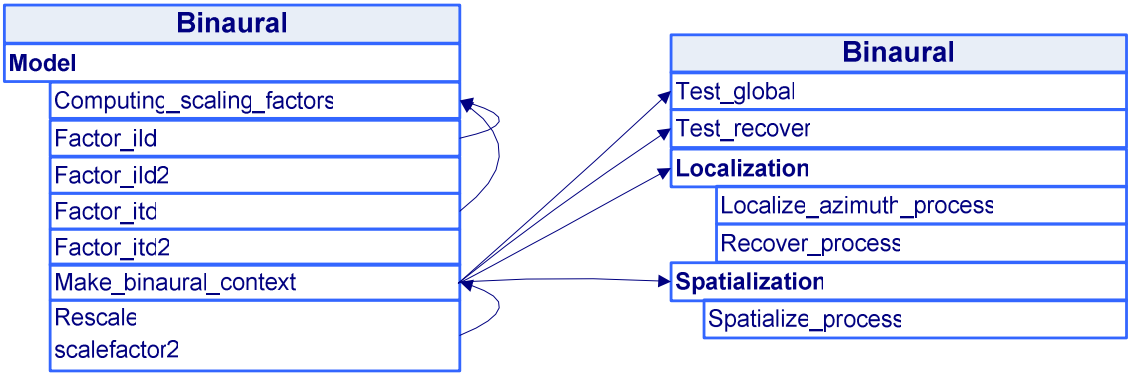


3. *Spatial* Folder

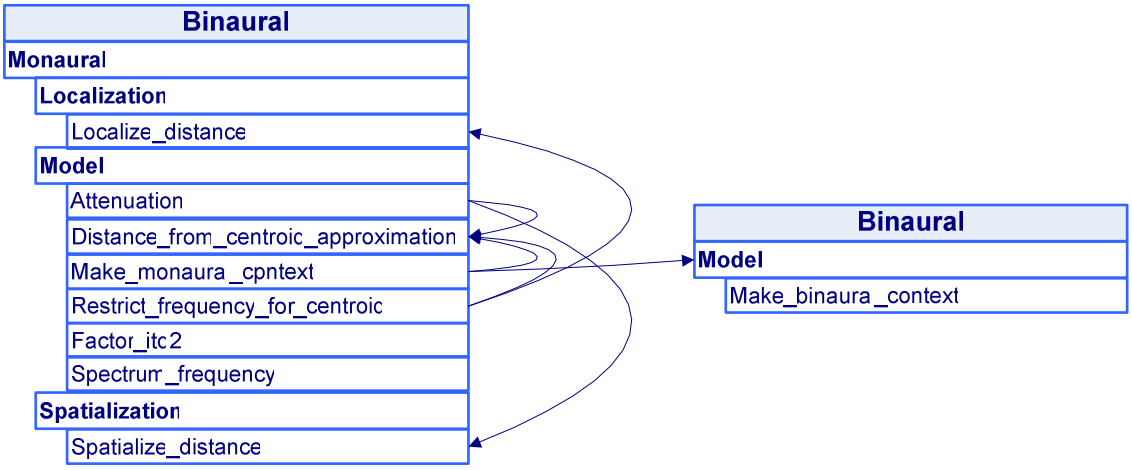
a. Localization



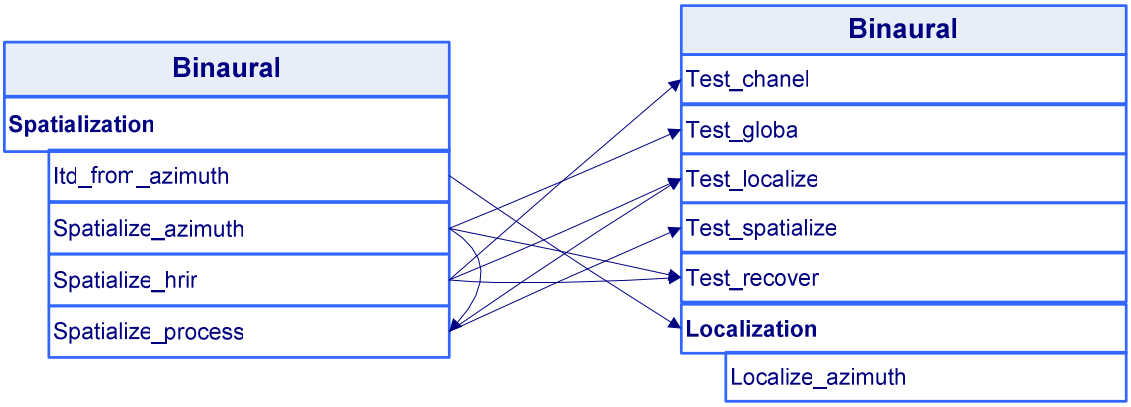
b. Model



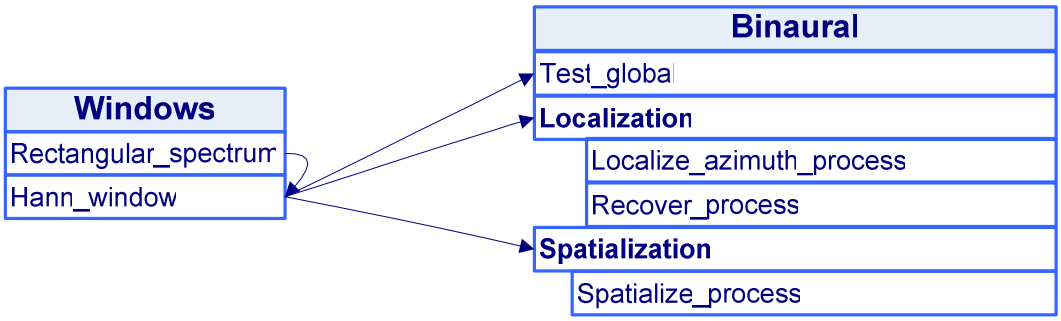
c. Monaural



d. Spatialization



4. *Windows* Folder



5. Zerophase Folder