The Hough Transform for Binaural Source Localization

Sylvain Marchand, Anne Vialard

▶ **To cite this version:**

Sylvain Marchand, Anne Vialard. The Hough Transform for Binaural Source Localization. Digital Audio Effects (DAFx09) Conference, Sep 2009, Como, Italy. pp.252–259. hal-00406275

HAL Id: hal-00406275
https://hal.science/hal-00406275

Submitted on 28 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THE HOUGH TRANSFORM FOR BINAURAL SOURCE LOCALIZATION

Sylvain Marchand
LaBRI – CNRS
University of Bordeaux 1
Talence, France
sylvain.marchand@labri.fr

Anne Vialard
LaBRI – CNRS
University of Bordeaux 1
Talence, France
anne.vialard@labri.fr

ABSTRACT

We introduce a new technique for the blind localization of several sound sources from two binaural signals. First, the binaural signals are organized as two-dimensional data where each sound source appears as a line. Second, the Hough transform is used to recognize these lines. The slopes of the lines give the mixing coefficients and directions of arrival (azimuths). Two variants of our technique are proposed, based on only one of the interaural level or time differences, respectively. Although a rapid comparison to a well-known localization method as well as promising results are shown, they are clearly not exhaustive and this paper should rather be regarded as a feasibility demonstration of the new technique.

1. INTRODUCTION

Sound source localization and separation is an active research topic in the audio community. The blind approach, with no information on the sound sources or the mixing process, is specially challenging. With this approach, one can only rely on perception, and a classic strategy is to use binaural cues [1, 2, 3, 4, 5, 6].

The Degenerate Unmixing and Estimation Technique (DUET) [1, 2] consists in building a 2-dimensional histogram from two interaural cues: the interaural level difference (ILD) and the interaural time difference (ITD). This 2D histogram can be regarded as an image, where sound sources appear as local maxima of intensity.

This approach has many drawbacks, since these cues are not independent and the time difference is estimated by the phase difference, which is ambiguous. The cues can be combined into a more efficient localization algorithm [3, 4], and / or used in a statistical approach [5, 6] leading to Gaussian mixture models.

In early attempts for sound localization, the signals at the two ears were used directly – instead of two binaural cues – to plot simple geometric structures (*e.g.* ellipses) from which the spatial information could be derived, at least in simple mixing situations.

In this paper, we show that it is possible to use these binaural signals and cues to build images in which each sound source appears as a line, even within a very general mixture model. The problem is then to recognize the lines in these images. This can be achieved by using the Hough transform [7], well known in image analysis. Thus, we propose to combine the interaural signals and cues to get the binaural data and Hough transforms to analyze their distributions.

To our knowledge, this paper is the very first attempt of such an "image + sound" combination. However, we must mention that Richard O. Duda has worked in both research areas: Hough transform for image analysis [8], and binaural sound analysis [9]. This article has to be regarded as our early experiments with the binaural Hough technique, and mainly as a demonstration of the feasibility of this technique. Extensive tests and comparison with other techniques, as well as the extension of the localization technique to source separation, are part of our future research.

The remainder of this article is organized as follows. Section 2 presents a very general mixture model, convolutive and with additive noise, giving the way the different sound sources combine at the two ears to form the binaural signals. Section 3 describes the binaural model showing how interaural cues can be used to organize the binaural data in linear structures, one line per source. Section 4 gives an overview of the Hough transform, and demonstrates its ability to recognize lines and stripes in our context. Then Section 5 shows how the combination of the binaural and Hough analyses, in the cases of interaural level or time differences, can lead to an interesting source localization technique. Finally, Section 6 reports our early experiments with this new localization technique and Section 7 concludes by giving directions for further research.

2. MIXTURE MODEL

We first consider a very general – quite realistic – mixture model.

2.1. Temporal Domain

The binaural signals $s_C(t)$ arriving at the two ears (C being the channel, L left or R right) are obtained by the addition of all the contributions of the M sound sources with a Gaussian white noise $n_C(t)$, each contribution $v_{m,C}(t)$ being the result of the convolution ($*$) of the m-th source signal $s_m(t)$ by the acoustic path $a_{m,C}(t)$ from the source to the ear. More precisely, we have:

$$s_C = \sum_{m=1}^{M} \underbrace{a_{m,C} * s_m}_{v_{m,C}} + n_C \quad (C \in \{L, R\}). \tag{1}$$

In the simple cases of a single sinusoid or a complex sound but monophonic and with scalar mixing coefficients, it is well-known that plotting s_L as a function of s_R gives remarkable geometric shapes (see Figure 1). However, in the complex case of several sinusoids or sources and general acoustic paths for the $a_{m,C}$ coefficients, the (s_R, s_L) plot is not tractable anymore.

2.2. Spectral Domain

Fortunately, a spectral representation can handle these cases (as shown by Figures 2 and 3). More precisely, we switch to the time-frequency plane by means of a short-time Fourier transform (STFT), the mixing equations being now:

$$S_C = \sum_{m=1}^{M} \underbrace{A_{m,C} \cdot S_m}_{V_{m,C}} + N_C \quad (C \in \{L, R\}) \tag{2}$$

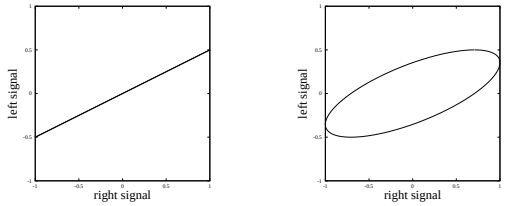

Figure 1: *Plotting (s_R, s_L) in the simple case of one ($M = 1$) sinusoid $s_1(t)$. Without phase difference among the mixing coefficients (left), the plot appears as a line which slope is the ratio between the left and right coefficients (here $1/2$). With an additional phase difference between the mixing coefficients (right), the plot appears as an ellipse: one axis is again directed by the amplitude of the ratio between the the coefficients, but the other is a function of the angle of this ratio (here $\pi/4$ rad).*

where $S(t, f)$ denotes the STFT of $s(t)$, and the convolution of Equation (1) is now a simple multiplication among spectra. Then, we take advantage of two properties of the spectral representation.

2.2.1. Orthogonality Assumption

First, the classic orthogonality assumption holds for most signals. More precisely, at a given point (t, f) of the time-frequency plane, there is a dominant source $m(t, f)$, the others being negligible, thus:

$$\forall t, \forall f, \quad S_C(t,f) \approx \underbrace{A_{m(t,f),C} \cdot S_{m(t,f)}}_{V_{m(t,f),C}} + N_C(t,f). \qquad (3)$$

2.2.2. Effect of Noise

Second, the additive noise and the sources are statistically independent, and thus the variances of the (zero-mean) signals sum up:

$$\forall t, \sum_f |S_C(t,f)|^2 = \sum_f |V_{m,C}(t,f)|^2 + \sum_f |N_C(t,f)|^2. \quad (4)$$

Moreover, in order to neglect the effect of the noise, we choose to consider only the local maxima of the magnitude spectra $|S_C|$, where the condition $|V_{m,C}|^2 >> |N_C|^2$ should hold (especially for low noise levels), and thus we have $|S_C|^2 \approx |V_{m,C}|^2$.

3. BINAURAL MODEL

Let us now investigate the relation between left and right signals:

$$V_{m,L}(t,f) = K_m(f) \cdot V_{m,R}(t,f) \qquad (5)$$

$$\text{where} \quad K_m = A_{m,L}/A_{m,R}. \qquad (6)$$

Before using perceptive considerations, with a model of the human head, let us first consider a simplified model, suitable for many signals produced by the audio industry.

3.1. Simple Panoramic Case

When using a mixing console, a gain and a delay are often applied to each channel of the mix, to produce an artificial spatialization.

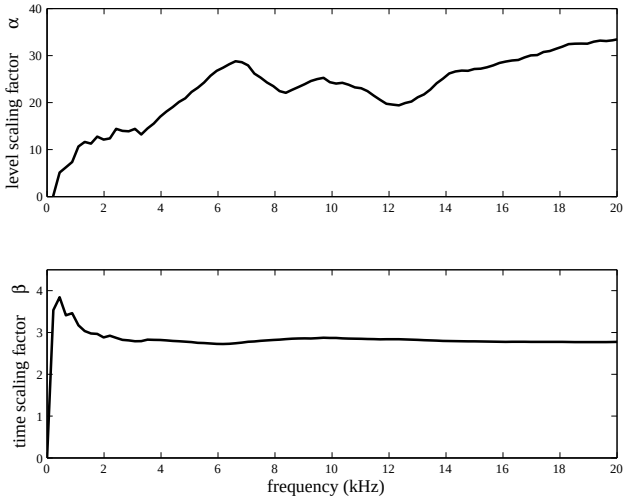

Figure 4: *Frequency-dependent scaling factors: α (top) and β (bottom).*

3.1.1. Level Difference (Gain)

Then $|K_m|$ is the mixing gain, the points $(x, y) = \left(|S_R|^2, |S_L|^2\right)$ (weighted by their power $|S|^2 = |S_L|^2 + |S_R|^2$) form lines of slopes $|K_m|^2$, as shown in Figure 2.

3.1.2. Time Difference (Delay)

Also $\angle K_m$ is a consequence of the time delay Δ_m, and more precisely $\angle K_m = 2\pi f \Delta_m$, thus the points $(x, y) = (f, \angle(S_L/S_R))$ form parallel stripes of slope $2\pi\Delta_m$, as shown in Figure 2. The stripes are due to the fact that the phase is measured modulo 2π.

3.2. Realistic Binaural Case

For natural signals, such as the ones recorded at the ears, K_m is not set by some mixing device, but rather a consequence of the propagation to the ears of the sound of the m-th source, positioned at some azimuth Θ_m.

3.2.1. Interaural Level Difference (ILD)

After Viste [3, 4], we have verified in [5] that

$$|K_m(f)| = 10^{\text{ILD}(\Theta_m, f)/20} \qquad (7)$$

$$\text{where} \quad \text{ILD}(\Theta, f) = \alpha(f)\sin(\Theta) \qquad (8)$$

and α is a scaling factor (see Figure 4) learned from the CIPIC database [9] (see [5, 10] for details). As a consequence, the points

$$(x, y) = \left(|S_R|^{20/\alpha}, |S_L|^{20/\alpha}\right) \qquad (9)$$

(again weighted by their power, this time estimated thanks to $|S|^2 = |S_L \cdot S_R|$ as indicated in [10]) should form a line of slope

$$a_m = 10^{\sin(\Theta_m)} \qquad (10)$$

which is a function of the azimuth Θ_m only. This is verified in practice, as shown in Figure 3.

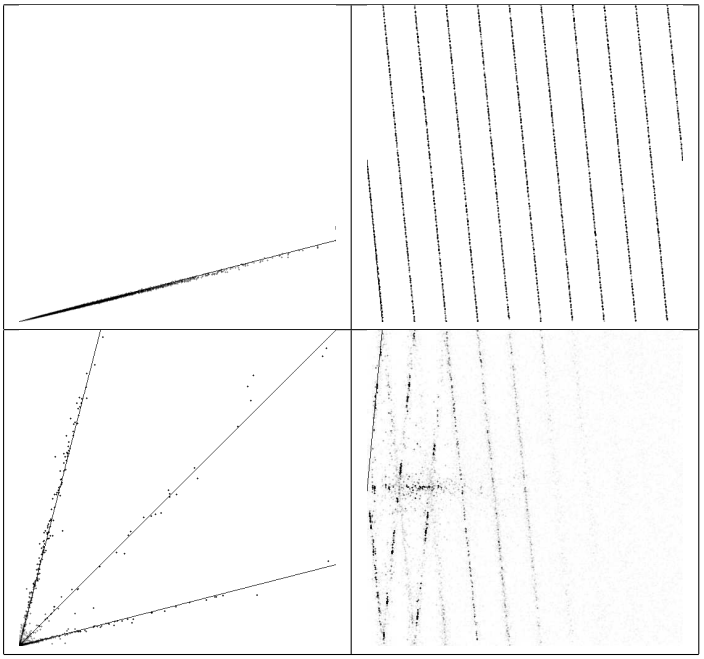

Figure 2: *Accumulation of the power of the (x, y) points (indicated with small crossed dots) according to $(x, y) = \left(|S_R|^2, |S_L|^2\right)$ (left) and $(x, y) = (f, \angle(S_L/S_R))$ (right), in the cases of a (monophonic) white noise mixed with a gain of $1/2$ and a delay of -20 samples (at $F_s = 44.1$kHz) (top) and 3 sources – singing voice with gain $1/2$ and delay -20 samples, xylophone with gain 1 and delay 0, and trumpet with gain 2 and delay $+20$ samples (bottom). One can clearly see one line per source for the gain case (left). For the delay case (right), one can notice series of parallel stripes, with a different slope for each source: negative, null (horizontal line), or positive. Solid lines indicate the slopes detected by the proposed localization method.*

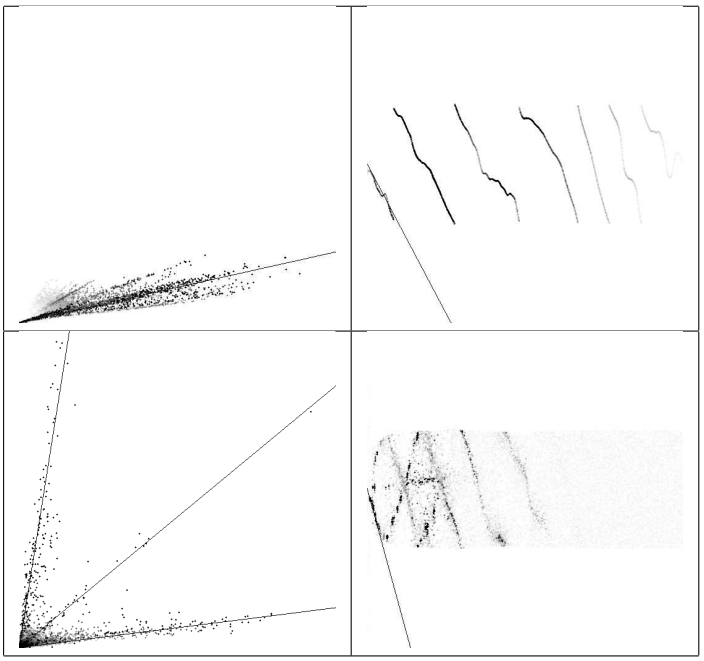

Figure 3: *Accumulation of the power of the (x, y) points (indicated with small crossed dots) according to Equations (9) (left) and (13) (right), in the cases of a (monophonic) white noise at azimuth -30° (top) and 3 sources – singing voice at -45°, xylophone at 0°, and trumpet at $+45^\circ$ (bottom). One can clearly see one line per source for the ILD case (left). For the ITD case (right), one can notice series of parallel stripes, with a different slope for each source: negative, null (almost horizontal line), or positive. The scaling of the vertical axis (which range is 2π) is a consequence of the scaling factor β in Equation (13). Solid lines indicate the slopes detected by the proposed localization method.*

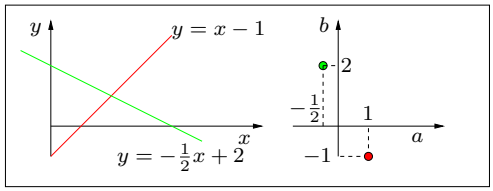

Figure 5: *Image space (left) and parameter space (right): Two representations of straight lines from their Cartesian equations.*

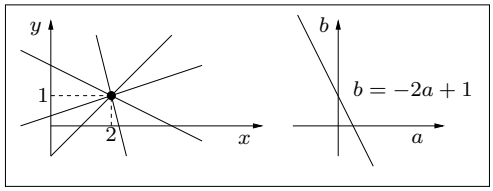

Figure 6: *The set of all the lines traversing one point (left) is represented by a unique line in the parameter space (right).*

3.2.2. Interaural Time Difference (ITD)

We have also shown in [10] that

$$\angle K_m(f) \equiv 2\pi\,\mathrm{ITD}(\Theta_m, f)\cdot f \mod 2\pi \quad (11)$$

$$\text{where} \quad \mathrm{ITD}(\Theta, f) = \beta(f)\,R\sin(\Theta)/c \quad (12)$$

and $R = 7.25$cm is the average head radius, $c = 335\text{m}\cdot\text{s}^{-1}$ is the sound celerity, and β is a scaling factor (see Figure 4) learned from the CIPIC database (see [9, 10]). As a consequence, the points

$$(x, y) = \left(f, \angle(S_L/S_R)^{1/\beta}\right) \quad (13)$$

should form parallel stripes of common slope

$$a_m = 2\pi\,R\sin(\Theta_m)/c \quad (14)$$

which is again a function of the azimuth Θ_m only. This is verified in practice, as shown in Figure 3.

We have also verified that panoramic signals generated with the simplified mixing process (Section 3.1) exhibits linear structures when analyzed with the general binaural model (Section 3.2).

4. HOUGH TRANSFORM

The Hough transform (HT) is an image analysis tool for the recognition of predefined shapes in an image (see for example [7] for an overview). We present here its use for the recognition of linear structures: straight lines (ILD case) and parallel stripes (ITD case).

4.1. Cartesian Version

As a preliminary, we explain how the HT can be used to recognize a line given by its Cartesian equation $y = ax + b$. Such a line can be represented by one point (a, b) in the *parameter space* which axes correspond respectively to the line slope a and to the y-intercept b. See Figure 5 for an illustration.

A point (x, y) in the image space can be represented by the line $b = -ax + y$ in the parameter space. In other words, a point is described by the set of all the lines passing through it. This is illustrated in Figure 6.

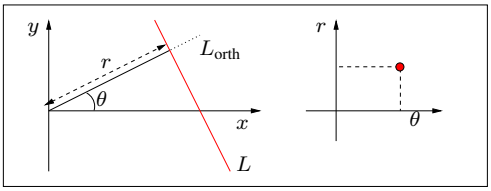

Figure 7: *Polar representation of a straight line in image space (left) and in parameter space (right).*

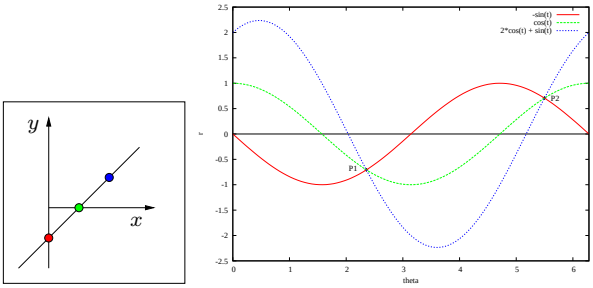

Figure 8: *Example of image space (three collinear points – left) versus parameter space (three sinusoids – right) for the polar representation.*

The main idea of the HT is that each image point votes for all the lines passing through it. A line is recognized if it gathers a significant number of votes. In practice, the Cartesian representation is not suitable (the distribution of the votes is hardly tractable, in particular vertical lines cannot be recognized) and a polar representation is used instead.

4.2. Polar Version

A straight line L is now given by the polar equation $r = x\cos(\theta) + y\sin(\theta)$. Let us denote by L_{orth} the perpendicular to L passing through the origin. The angle θ is the angle from the x-axis to L_{orth}. The value r is the distance between L and the origin. These definitions are illustrated in Figure 7.

The set of all the lines traversing a point (x, y) is represented by the sinusoid

$$r = x\cos(\theta) + y\sin(\theta) \quad (15)$$

in the parameter space. In practice, we efficiently generate the sinusoid using one cosine function, since

$$x\cos(\theta) + y\sin(\theta) = a\cos(\theta + \phi) \quad (16)$$

$$\text{where} \quad a = |z|, \phi = \angle z, z = x - iy$$

that can be computed using incremental methods. In the example of Figure 8, the three collinear points $(0, -1)$, $(1, 0)$ and $(2, 1)$ are respectively represented by the sinusoids $r = -\sin(\theta)$, $r = \cos(\theta)$ and $r = 2\cos(\theta) + \sin(\theta)$. The three curves intersect at points $P_1 = (\theta = 3\pi/4, r = -\sqrt{2}/2)$ and $P_2 = (\theta = 7\pi/4, r = \sqrt{2}/2)$.

Remark that P_1 and P_2 correspond to the same line of Cartesian equation $y = x - 1$. In fact, when considering both positive and negative values for r, a range of π rad is sufficient for θ to avoid redundancy in the transform (indeed, considering $\theta + \pi$ instead of θ in Equation (15) is equivalent to considering θ but with $-r$ instead of r). We will use the $[-\pi/2, +\pi/2]$ range.

For computation, the parameter space is digitized as a numeric matrix with θ in row and r in column. For each considered point in the image space, its sinusoid is traced on the matrix by incrementing each encountered cell. In practice, to allow for some errors and compensate for the discrete nature of the Hough matrix, we accumulate a Gaussian distribution and thus the neighboring cells are affected as well. When all the points are processed (all the "votes" are taken into account), the maxima in the accumulator matrix give the most salient straight lines in the image.

From the (r, θ) coordinates of these maxima, one can recover the slope a and y-intersect b of the lines since

$$a = -1/\tan(\theta) \quad \text{and} \quad b = r/\sin(\theta). \tag{17}$$

4.3. Line Recognition (1D Transform)

When using ILDs (see Figures 11 and 14), we are interested in recognizing lines passing through the origin of the image, that is with $r = 0$. Each (x, y) point contributes to the transform at

$$\theta = -\arctan(x/y) \tag{18}$$

and thus designing an efficient implementation of this unidimensional transform (function of θ only) is easy. Although θ ranges from $-\pi/2$ to $+\pi/2$, only the first half of the range is useful, since it corresponds to the positive slopes.

4.4. Stripes Recognition (2D Transform)

When dealing with ITDs (see Figures 12 and 15), then the r dimension is useful. We are again interested in recognizing a line passing through the origin, but since this line is produced by a wrapped phase it is in fact a series of parallel stripes. These stripes are uniformly spaced, and cause in the full (two-dimensional) Hough transform, at the corresponding θ, a periodicity in r with a period

$$\Delta_r = 2\pi|\sin(\theta)| \tag{19}$$

that can be observed in Figures 9, 12, and 15. As seen on these figures, another way of observing this periodicity is to consider the Fourier transform (FT) of each column (θ dimension) of the Hough transform (HT). The local maxima of the combined HT/FT transform should be located at the θ corresponding to the sound sources. This is verified in practice, at least in monophonic cases.

5. SOURCE LOCALIZATION

We now present several localization techniques: a classic technique based on a power histogram, and the new proposed method using the Hough transform. In each case, either the interaural level differences (ILDs) or the interaural time differences (ITDs) can be used, leading to two sub-techniques. Also, the new approach works for panoramic and binaural mixes (see Section 3).

5.1. ILD-Based Localization

One can estimate the positions of the sources using only the ILDs. The problem with the ILDs is that they are quite dispersed, and cause a bias towards the extreme azimuths. Moreover, they are problematic at low frequencies, and hardly tractable in practice when dealing with reverberant conditions [11].

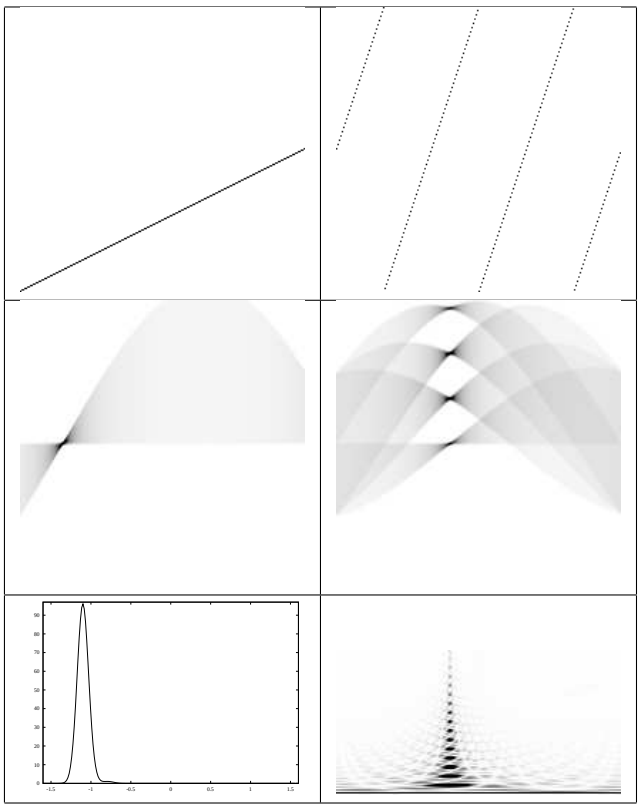

Figure 9: *The Hough transform for the recognition of linear structures: straight lines (left) or parallel stripes (right). In the first case (left), the origin of the image (top) is at its bottom-left corner. Since the line passes though this origin, the 2D Hough transform (middle) exhibits a maximum for $r = 0$ (center of the vertical axis of the HT image) at the θ (horizontal axis of the HT image) corresponding to the slope of the line. Then the 1D version of the Hough transform (bottom) clearly shows this maximum at the same value of θ. In the second case (right), the origin of the image (top) is still at the left of the horizontal axis (frequency), but now at the center of the vertical axis (phase). The Hough transform (middle) exhibits a series of local maxima (one per line segment), in a r-periodical way for the θ corresponding to the slope of the lines. Finally, the Fourier transform (bottom) for each θ shows a maximum at the same abscissa.*

5.1.1. Classic Analysis

The existing approach [4, 5] consists in accumulating the power of each time-frequency point (t, f) at the corresponding azimuth $\Theta(t, f)$ in some histogram. First, the ILD is estimated with

$$\hat{\text{ILD}}(t, f) = 20 \log_{10} \left| \frac{S_L}{S_R}(t, f) \right| \tag{20}$$

then Θ is obtained thanks to the inverse of the model Equation (8):

$$\Theta_L(t, f) = \arcsin \left(\frac{\hat{\text{ILD}}(t, f)}{\alpha(f)} \right). \tag{21}$$

This way, a power histogram is built and its local maxima should correspond to the directions of arrival of the sound sources. An example with 3 sources is illustrated in Figure 10.

5.1.2. Hough Analysis

The proposed approach consists in first building the 1D Hough transform (see Section 4.3) of the (x, y) points given by Equation (9), then searching for the local maxima in this transform. As shown in Figures 11 and 14, these maxima are located at the θ_m corresponding to the slopes a_m (see Equation (17)) of the linear distributions formed by the points, and from these slopes it is possible to find in turn the azimuths Θ_m of the sound sources by inverting Equation (10).

Remark that the Hough transform is particularly well-suited here, since it gathers the information from independent data points, such as the spectral atoms of the STFT, also possibly coming from different sound sources.

5.2. ITD-Based Localization

The ITDs are more accurate, and more resistant to reverberation. The problem is that they are ambiguous, because they are derived from the interaural phase difference, measured modulo 2π.

5.2.1. Classic Analysis

The existing approach [4, 10] consists again in accumulating in some histogram the power of each time-frequency point (t, f) at the corresponding azimuth $\Theta(t, f)$. This time, the ILD and ITD informations are combined to yield a better azimuth estimate. First, the ILD is estimated using Equation (20). Second, the ITD information is estimated with

$$\hat{\text{ITD}}_p(t, f) = \frac{1}{2\pi f} \angle \left(\frac{S_L}{S_R}(t, f) \right) + 2\pi p. \tag{22}$$

The coefficient p highlights the fact 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. To obtain an estimation of the azimuth candidate for each p, we invert Equation (12):

$$\Theta_{T,p}(t, f) = \arcsin \left(\frac{c \cdot \hat{\text{ITD}}_p(t, f)}{R \cdot \beta(f)} \right). \tag{23}$$

The $\Theta_L(t, f)$ estimates are more dispersed, but never ambiguous, 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 time-frequency point, since it exhibits a smaller deviation:

$$\Theta(t, f) = \Theta_{T,q}(t, f) \tag{24}$$

$$\text{where} \quad q = \text{argmin}_p \left| \Theta_L(t, f) - \Theta_{T,p}(t, f) \right|.$$

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

$$p_r = \left(f \cdot \text{ITD}(\Theta_L, f) - \frac{1}{2\pi} \angle \frac{S_L(t, f)}{S_R(t, f)} \right). \tag{25}$$

This way, a more accurate power histogram is built and its local maxima should correspond to the directions of arrival of the sound sources. An example with 3 sources is illustrated on Figure 13.

5.2.2. Hough Analysis

Here, we propose a new approach using only the ITD information (whereas the classic method requires also the ILD information). The Hough transform (see Section 4.4) is performed on the (x, y) points obtained thanks to Equation (13), then a Fourier transform is computed in turn and its local maxima should correspond to sound sources. From the angle θ of each maximum, a slope a is computed using Equation (17). In the panoramic case, the time delay is given by $\Delta_m = a_m/(F_s/2)$, where F_s denotes the sampling frequency. In the binaural case, the azimuth Θ_m is estimated from the slope a_m by inverting Equation (14).

Remark that the Fourier transform requires a certain number of periodicities. For extreme azimuths, there can be up to 20 of them (because the ITD is then close to 1ms, and see Equation (11)). But they may be insufficient with azimuth close to 0. If needed, this number can be artificially increased by elevating the S_L/S_R ratio to the power of some positive integer P. It increases the wrapping effect of the phase ratio $\angle(S_L/S_R)$, multiplies the number of stripes by P, but also their slope by the same factor.

6. EXPERIMENTS

In our early experiments, we considered four tests signals (monophonic or polyphonic examples obtained with either panoramic or binaural mixing processes), then we estimated the mixing parameters / azimuths using either the ILD or the ITD information.

6.1. Test Signals

For the monophonic cases ($M = 1$), we use a Gaussian white noise (broadband signal). For the polyphonic cases, we consider $M = 3$ sound sources: the singing voice of Suzanne Vega (the famous excerpt from "Tom's Dinner") at the right, a xylophone in the middle, and a trumpet at the left. All sound sources are sampled at $F_s = 44.1$kHz, zero-mean, and normalized to the same duration and level prior to the mixing process. In the polyphonic cases, we use additive noises n_C of level -20dB (see Equation (1)). The mixing parameters or locations of the sources are indicated in Tables 1 or 2, respectively. For the binaural case, the sources are spatialized using the head-related impulses responses (HRIRs) of the KEMAR manikin [12], found in the CIPIC database [9].

6.2. Experimental Results

These test signals were analyzed using the classic and proposed methods, with or without the ITD information. For the STFT, we used frames of $N = 2048$ samples and the Hann window.

sound source	amplification gain		time delay in samples	
white noise	1/2	(0.51)	-20	(-21)
singing voice	1/2	(0.51)	-20	(N/A)
xylophone	1	(1.00)	0	(N/A)
trumpet	2	(1.97)	+20	(+21)

Table 1: *Mixing coefficients (and their estimations in parentheses, if available) for the panoramic case.*

sound source	azimuth in degrees		
white noise	-30	(-40.59 [-37],	-23.32 [-24])
singing voice	-45	(-63.96 [-52],	**-50.00 [-36]**)
xylophone	0	(-4.72 [-2],	N/A [+3])
trumpet	+45	(**+53.16 [+65]**,	N/A [+45])

Table 2: *Source locations (and their estimations with ILDs, then with ITDs) for the binaural case. [The reference estimations using the classic method are indicated in square brackets.]*

For the panoramic mixes, the results are shown in Figures 2, 11, and 12, and summarized in Table 1. In all cases the linear structures appear in the data, the Hough transform is able to recognize them, and the mixing coefficients are estimated with a great accuracy. The only problem is with the estimation of multiple time delays in the polyphonic case, since the Fourier transform of the Hough transform is too messy (see Figure 12) and the estimation is done only for one source – corresponding to the global maximum.

For the binaural mixes, the results are shown in Figures 10 and 13 (classic method), Figures 3, 14, and 15 (proposed method), and summarized in Table 2. Again, in all cases the linear structures appear in the data, the Hough transform is able to recognize them, and the directions of arrival are estimated with a sufficient accuracy. As mentioned in Section 5, the ILD-based method overestimates the azimuths, whereas the ITD-based method shows a better accuracy. The proposed method could even be a serious challenger for the state-of-the-art method (see the results in bold in Table 2). Unfortunately, the Fourier transform of the Hough transform is again messy (see Figure 15) and thus for now the estimation is done only for one source – again using the global maximum.

7. CONCLUSION AND FUTURE WORK

We have in fact introduced two new techniques for binaural source localization, using the Hough transform. The first one takes advantage of the ILD information only, uses a 1D Hough transform, and is roughly equivalent to the existing histogram-based method – with the same drawbacks (overestimation of the azimuths, poor resistance to reverberation). The second one takes advantage of the ITD information only, uses a 2D Hough transform, and shows a lower estimation bias. Unlike the existing histogram-based method, the ILD information is not used anymore, which could be a great advantage especially in reverberant conditions. Unfortunately, the interpretation of the Hough transform is more complicated and for now we localize only the dominant source. To be able to localize all the sources, we need to compute the accumulation of the votes for the multiples of Δ_r (see Section 4.4). For now this is done using a Fourier transform. However, we need a greater resolution for low values of θ. Moreover, computing first the (discrete) Hough transform then the (discrete) Fourier transform is not optimal. For these reasons, we are now working on the design of a specific transform.

8. ACKNOWLEDGMENTS

This research was partly supported by the French GIP ANR, DESAM *(Décomposition en Éléments Sonores et Applications Musicales)* project (ANR-06-JCJC-0027-01).

9. REFERENCES

[1] Alexander Jourjine, Scott Rickard, and Özgür Yılmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing n Sources From 2 Mixtures," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, pp. 2985–2988.

[2] Özgür Yılmaz and Scott Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] Harald Viste, *Binaural Localization and Separation Techniques*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2004.

[4] Harald Viste and Gianpaolo Evangelista, "Binaural Source Localization," in *Proceedings of the Digital Audio Effects (DAFx) Conference*, Naples, Italy, October 2004, Federico II University, pp. 145–150.

[5] Joan Mouba and Sylvain Marchand, "A Source Localization/Separation/Respatialization System Based on Unsupervised Classification of Interaural Cues," in *Proceedings of the Digital Audio Effects (DAFx) Conference*, Montreal, Quebec, Canada, September 2006, McGill University, pp. 233–238.

[6] Michael I. Mandel and Daniel P. W. Ellis, "EM Localization and Separation Using Interaural Level and Phase Cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2007, pp. 275–278.

[7] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, Prentice-Hall, 2nd edition, 2002.

[8] Richard O. Duda and Peter E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Communications of the Association for Computing Machinery (ACM)*, vol. 15, pp. 11–15, January 1972.

[9] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos 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, USA, October 2001, pp. 99–102.

[10] Joan Mouba, Sylvain Marchand, Boris Mansencal, and Jean-Michel Rivet, "RetroSpat: a Perception-Based System for Semi-Automatic Diffusion of Acousmatic Music," in *Proceedings of the Sound and Music Computing (SMC) Conference*, Berlin, Germany, July/August 2008, pp. 33–40.

[11] Barbara G. Shinn-Cunningham, Norbert Kopco, and Tara J. Martin, "Localizing Nearby Sound Sources in a Classroom: Binaural Room Impulse Responses," *Journal of the Acoustical Society of America (ASA)*, vol. 117, no. 5, pp. 3100–3115, May 2005.

[12] William G. Gardner and Keith Martain, "HRTF Measurements of a KEMAR Dummy-Head Microphone," Tech. Rep., MIT Media Lab, May 1994.

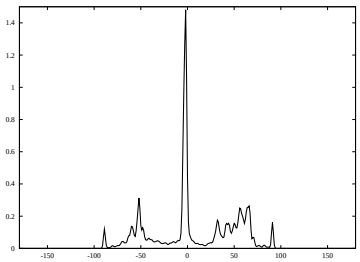

Figure 10: *Power histogram obtained using only the* **ILD** *information, for a mix of* 3 *sources located at azimuths* $-45°$, $0°$, *and* $+45°$. *The central source is clearly visible. However, the energy of the other sources is spread towards the extreme azimuths* ($\pm 90°$), *where spurious peaks appear. The azimuth estimation gets biased.*

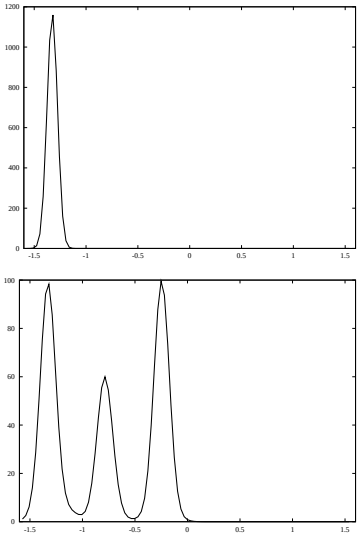

Figure 11: *1D Hough transforms obtained for* **panoramic** *signals made of* 1 *(top) or* 3 *(bottom) source(s) – see Sections 5.1 and 6 for details.*

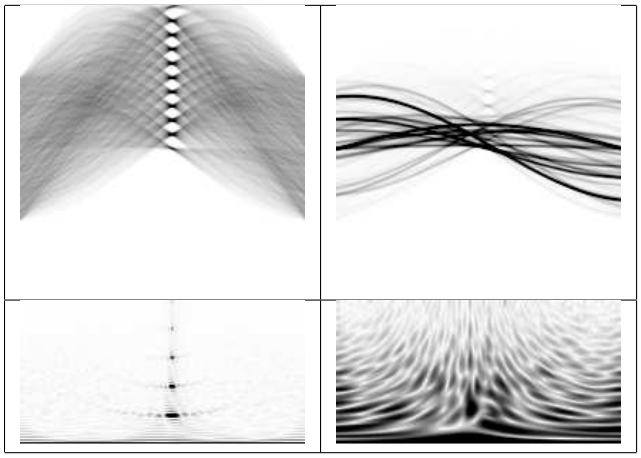

Figure 12: *2D Hough transforms (top) and associated Fourier transforms (bottom), obtained for* **panoramic** *signals made of* 1 *(left) or* 3 *(right) source(s) – see Sections 5.2 and 6 for details.*

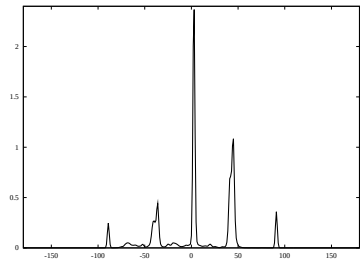

Figure 13: *Power histogram obtained using the full* **ILD+ITD** *information, for a mix of* 3 *sources located at azimuths* $-45°$, $0°$, *and* $+45°$. *The sources are clearly visible. However, spurious peaks again show up at extreme azimuths* ($\pm 90°$). *These peaks are consequences of errors of the ILD model mostly at low frequencies.*

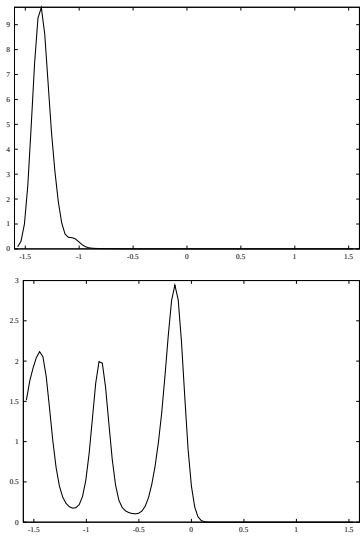

Figure 14: *1D Hough transforms obtained for* **binaural** *signals made of* 1 *(top) or* 3 *(bottom) source(s) – see Sections 5.1 and 6 for details.*

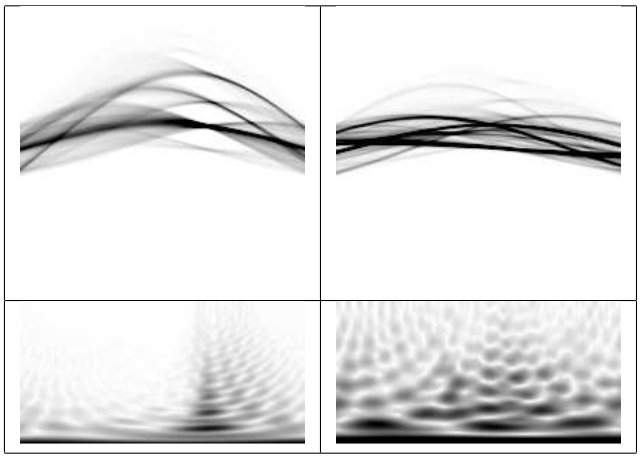

Figure 15: *2D Hough transforms (top) and associated Fourier transforms (bottom), obtained for* **binaural** *signals made of* 1 *(left) or* 3 *(right) source(s) – see Sections 5.2 and 6 for details.*