PERFORMANCE OF SOURCE SPATIALIZATION AND SOURCE LOCALIZATION ALGORITHMS USING CONJOINT MODELS OF INTERAURAL LEVEL AND TIME CUES

Joan Mouba

SCRIME – LaBRI, University of Bordeaux 1
351 cours de la Libération, F-33405 Talence cedex, France
joan.mouba@labri.fr

ABSTRACT

In this paper, we describe a head-model based on interaural cues (*e.g.* interaural level differences and interaural time differences). Based on this model, we proposed, in previous works, a binaural source spatialization method (SSPA), that we extended to a multi-speaker spatialization technique that works on a speaker array in a pairwise motion (MSPA) [1], [2]. Here, we evaluate the spatialization techniques, and compare them to well-known methods (*e.g.* VBAP (Vector Base Amplitude Panning) [3]). We also test the robustness of a adapted conjoint localization method under noisy and reverberant conditions; this method uses spectra of recorded binaural signals, and tries to minimize the distance between the ILD and ITD based azimuth estimates. We show comparative results with the PHAT generalized cross-correlation localization method [4].

1. INTRODUCTION

In active listening applications, the spatialization and the localization are very important tasks. The spatialization allows the projection of a source in the space surrounding the listener, while the localization is the reciprocal operation, that consists in finding the source position. An overview of spatialization and localization techniques is given in [5].

Here, we considered punctual and omni-directional sound sources in the horizontal plane where both the listeners and the speakers are on the same ground. Each source is located by its (ρ, θ) coordinates, where ρ is the distance of the source to the listener head's center and θ azimuth angle.

In a binaural context, the difference in amplitude or Interaural Level Difference (ILD, expressed in decibels – dB) and in arrival time or Interaural Time Difference (ITD, expressed in seconds) are the main spatial cues for the auditory system [6]. In fact, a sound source positioned towards 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 [7].

We show the usefulness of the parametric ITD model from which we derive a binaural spatialization algorithm (SSPA: Source SPAtialization), and we extended this method to a multi-speaker system (MSPA: Multi-diffusion SPAtialization). The MSPA technique operates on loudspeakers in a pairwise manner. The computation of the panning coefficients are based on a adaptation of a static matrix of Head-Related Transfer Functions (HRTFs), which leads to frequency-dependent complex coefficients. We compare the MSPA to the classical VBAP, which also works on a pairwise manner but uses frequency independent panning coefficients. We also demonstrate the competence of our adaptation of the conjoint localization method after Viste [8]. This method uses conjointly the azimuth estimation from ILD and ITD to derive a robust localization for low and high frequencies [2].

This paper is structured as follows. First, we present the binaural model in Section 2. The associated binaural spatialization and multi-speaker spatialization techniques are detailed in Section 3. The conjoint localization method is explained in Section 4. The Section 5 is reserved for the analysis of the experiments results.

2. HEAD MODEL

2.1. Stereo model

A (vibrating) sound source s radiates acoustic waves, that will reach the left (L) and right (R) ears through different acoustic paths, characterized with a pair of filters, called Head-Related Impulse Responses (HRIRs). HRIRs are subject-dependent. The CIPIC database [9] contains samples for different listeners and different directions of arrival.

For a source s located at the azimuth θ, the left (x_L) and right (x_R) signals are given by:

$$x_L = s * \text{HRIR}_L(\theta), \quad (1)$$

$$x_R = s * \text{HRIR}_R(\theta), \quad (2)$$

where $*$ denotes the convolution among time-domain signals. HRIR characterizes generally anechoic environments. In a room, the HRIRs are replaced by BRIRs (Binaural Room Impulse Responses).

2.2. Interaural Level Differences

Viste [8] expressed the ILDs as functions of $\sin(\theta)$, with:

$$\text{ILD}(\theta, f) = \alpha(f) \sin(\theta), \quad (3)$$

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

Practically, the ILD for each time-frequency bin is measured by the ratio of the left (X_L) and right(X_R) short-time spectra with:

$$\text{ILD}(t, f) = 20 \log_{10} \left| \frac{X_L(t, f)}{X_R(t, f)} \right|. \quad (4)$$

2.3. Interaural Time Differences

After Woodworth [10], Viste [8] proposed a ITD model based on $\sin(\theta) + \theta$. However, from the theory of the diffraction of an harmonic plane wave by a sphere (the head) [11], we proposed a ITD

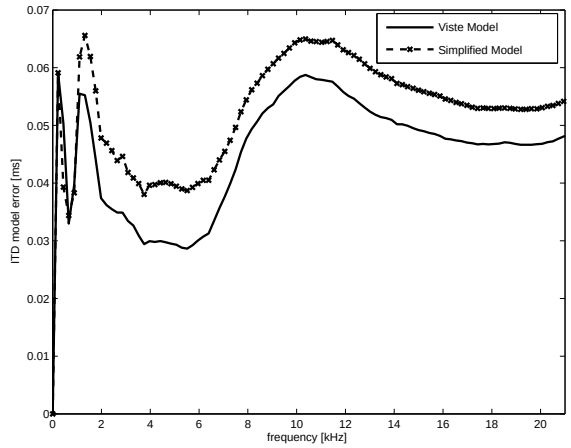


Figure 1: *Average ITD model error over the CIPIC Database.*

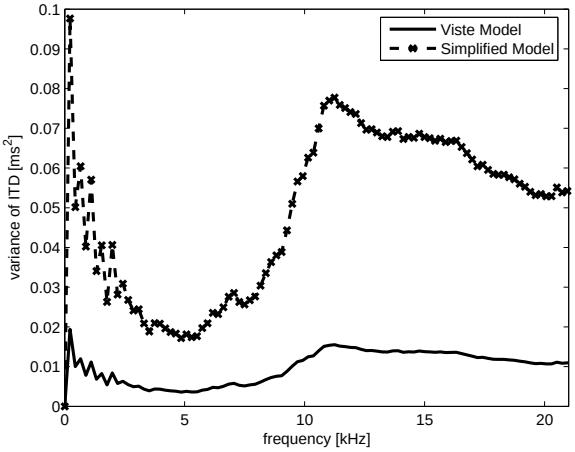


Figure 2: *Inter-subject variance over the CIPIC Database.*

model proportional to $\sin(\theta)$ [2]. These ITD models are given by:

$$\text{ITD}(\theta, f) = \beta(f) r(\sin(\theta) + \theta)/c, \quad (5)$$
$$\text{ITD}(\theta, f) = \gamma(f) r \sin(\theta)/c, \quad (6)$$

where β and γ are the average scaling factors that best suits the Viste and our sinusoidal model, in the least-square sense, for each listener of the CIPIC database, r denotes the head radius, and c is the sound celerity. The overall error of our model over the CIPIC database is 0.052 ms, which is comparable to the 0.045 ms error of the Viste's model. The average model errors and inter-subject variances of both models are depicted in Figures 1 and 2.

In practice, our model is easily invertible, which is suitable for sound localization, while the inversion of the $\sin(\theta) + \theta$ model by Viste requires more complex computations, and introduced mathematical approximation errors at the lateral azimuths (see [2]). In the next sections, we consider the $\sin(\theta)$ based model for the ITD.

Given the short-time spectra of the left (X_L) and right (X_R) channels, the ITD for each time-frequency bin is measured with:

$$\text{ITD}_p(t, f) = \frac{1}{2\pi f}\left(\angle \frac{X_L(t,f)}{X_R(t,f)} + 2\pi p\right). \quad (7)$$

The coefficient p highlights 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.

3. SPATIALIZATION TECHNIQUES

3.1. Binaural Spatialization

We proposed the SSPA binaural spatialization technique for headphones listening conditions. In this case, each ear receive only the sound from one earphone. Thus the encoded spatial cues are not affected by any cross-talk signals between earphone speakers.

The SSPA relies on the symmetry among the left and the right ears. 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 (3) and (6), and:

$$X_L(t, f) = X(t, f) \cdot 10^{+\Delta_a(f)/2} e^{+j\Delta_\phi(f)/2}, \quad (8)$$
$$X_R(t, f) = X(t, f) \cdot 10^{-\Delta_a(f)/2} e^{-j\Delta_\phi(f)/2}, \quad (9)$$

where Δ_a and Δ_ϕ are given by:

$$\Delta_a(f) = \text{ILD}(\theta, f)/20, \quad (10)$$
$$\Delta_\phi(f) = \text{ITD}(\theta, f) \cdot 2\pi f. \quad (11)$$

The conjoint control of amplitude and phase should provide better audio quality than amplitude-only spatialization. Although, errors on phase could deteriorate the overall audio quality [12]. We reach a remarkable spatialization realism through informal listening tests with AKG K240 studio headphones.

3.2. Multi-speaker Spatialization

We proposed the MSPA which is a extension of the SSPA technique to a multi-source multi-speaker system. In a setup with more than 2 speakers, the system adapts to different speaker configuration through the classic pairwise paradigm [13] in a stereophonic display. It consists in choosing for a given target source only the two speakers closest to it (in azimuth): one at the left of the source, the other at its right. In this case, the sound from each loudspeaker is heard by both ears. Thus, the stereo sound is filtered by a matrix of four transfer functions ($C_{ij}(f, \theta)$) between loudspeaker j and ear i ($i, j = L, R$) [1]. 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 (8) and (9)) are then given by:

$$K_L(t, f) = C \cdot (C_{RR} H_L - C_{LR} H_R), \quad (12)$$
$$K_R(t, f) = C \cdot (-C_{RL} H_L + C_{LL} H_R) \quad (13)$$

with the determinant computed as:

$$C = 1/(C_{LL} C_{RR} - C_{RL} C_{LR}). \quad (14)$$

For the stability of the solutions, the implementation must handle especially the cases where $|C| = 0$ (or close to zero) at any frequency.

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 the panning coefficients K_L and K_R, respectively:

$$Y_L(t,f) = K_L(t,f) \cdot X(t,f), \quad (15)$$
$$Y_R(t,f) = K_R(t,f) \cdot X(t,f). \quad (16)$$

4. SOURCE LOCALIZATION

4.1. Generalized cross-correlation source localization

Many source localization algorithms exist in the literature [14]. A useful method known to be robust in noisy and reverberant conditions is the PHAT-GCC method (or Generalized Cross-Correlation with Phase Transform) [4]. It consists in computing the inverse Fourier transform of a pre-filtering cross-power spectrum with:

$$G_{LR}(\tau) = \int_{-\infty}^{\infty} \frac{X_L(t,f)X_R^*(t,f)}{|X_L(t,f)X_R^*(t,f)|} e^{j2\pi f\tau} df, \quad (17)$$

where * denotes the complex conjugate operation, τ the time difference between the left and the right channels. The weighting functions allows to consider a finite signal length. Moreover, interferences are easily detect in the frequency domain. By dividing the cross-power by its magnitude, the PHAT function ensures a constant energy over all frequencies. Thus, when no single frequency dominates, the effect of reverberation is canceled out when averaged over many frequencies. We may observe local maxima in the result correlation function. The dominant peak is detected as the right DOA estimation. Though care must be taken for frequency points with near zero amplitude. The interaural ITD is given by:

$$\text{ITD} = \text{argmax}_\tau \left| G_{LR}(\tau) \right|. \quad (18)$$

The best mapping from the ITD to the azimuth was obtained by Equation 6 with the frequency independent factor $\xi(f) = 2.5$.

4.2. Conjoint source localization

In Auditory Scene Analysis (ASA), ILDs and ITDs are the most important cues for source localization. Lord Rayleigh mentioned in his Duplex Theory [7] 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 (Equation (4)) is just a matter of inverting Equation (3):

$$\theta_L(t,f) = \arcsin\left(\frac{\text{ILD}(t,f)}{\alpha(f)}\right). \quad (19)$$

Similarly, using the ITD information (see Equation (7)), to obtain an estimation of the azimuth candidate for each p, we invert Equation (6):

$$\theta_{T,p}(t,f) = \arcsin\left(\frac{c \cdot \text{ITD}_p(t,f)}{r \cdot \beta(f)}\right). \quad (20)$$

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:

$$\theta(t,f) = \theta_{T,m}(t,f), \quad (21)$$

with $m = \text{argmin}_p \left| \theta_L(t,f) - \theta_{T,p}(t,f) \right|$. Practically, the choice of p can be limited among two values ($\lceil p_r \rceil$, $\lfloor p_r \rfloor$), where

$$p_r = \left(f \cdot \text{ITD}(\theta_L, f) - \frac{1}{2\pi} \angle \frac{X_L(t,f)}{X_R(t,f)} \right). \quad (22)$$

For each frequency bin of each discrete spectrum, an azimuth is estimated and the corresponding power is accumulated in the histogram at this azimuth. An estimate of the azimuth of the source can be obtained as the peak in the built energy histogram (see [2]).

Figure 3 depicts the power histogram of a mixture of two speech signals at -30° and $+30^\circ$. The mixed binaural signals were produced by convolution of mono sources with the HRIRs of the KEMAR mannequin (see [9]). The histogram is enhanced by a polynomial smoothing operator $h_s(\theta)$ and then thresholded $h_t(\theta)$. A spurious source remains about azimuth -45°. Here, the threshold level is set as a fractional of the maxima of the histogram. In our experiments, we obtain appreciable results with a threshold set to the third of the maxima: $threshold = \frac{1}{3} max(h(\theta))$. Then, the number of peaks is a good estimator of the mixture's order.

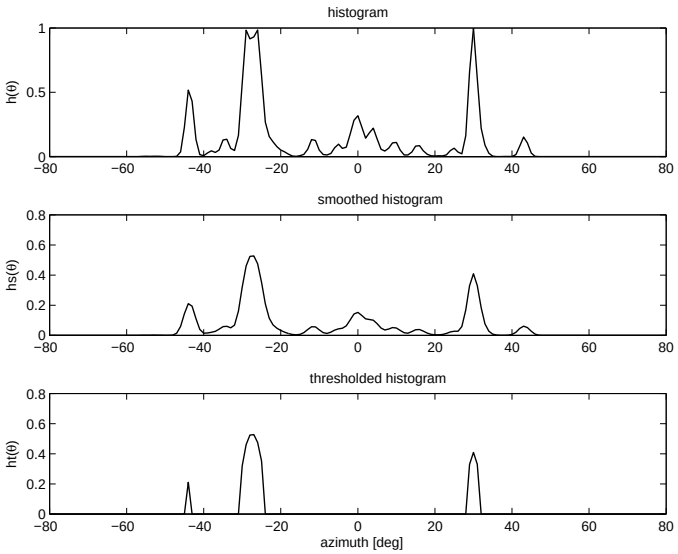


Figure 3: *Mixture of two sources at* $(-30^\circ, 30^\circ)$

.

From top to bottom: original histogram $h(\theta)$, smoothed histogram $h_s(\theta)$ and histogram after thresholding $h_t(\theta)$. The peak number decreases from $K = 13$ to $K = 3$, only one spurious peak remains.

5. PERFORMANCE ANALYSIS

5.1. Spatialization results

We conducted objective and subjective tests to evaluate the spatial realism and the sound quality of the proposed SSPA and MSPA

methods. The spatial realism describes the subjective accuracy of the projection in space, and the sound quality is related to the overall perceptual sensation (frequency content, loudness, listening pleasure).

5.2. SSPA performance

5.2.1. Subjective tests

For the subjective tests [1], we had 10 subjects, all members of the sound processing Team, familiar with sound evaluation. First, the suject had to judge the quality between the original sound and its spatialized version. We use a 5 points scale (1. perfect, 2: minor artifact, 3: distorded but intelligible, 4: very distorded and 5: not intelligible). The methods, SSPA and SHRIR (Spatialization with Head-Related Impulse Responses) have a average rate about 2, with a little preference for the SSPA method [2].

Second, we compare SSPA signals at different locations. We notice no confusion between left and right. For a resolution of 5°, 90% of the subjects could not differ the relative localization between two consecutive position. For cross-pair (one from SHRIR and one from SSPA), for the same position, about 15% perceived the SSPA sound more lateralized than his concurrent, the rests detect the same location. This highlights that our head model match heads of a large number, and does not distord the real location.

5.2.2. Objective tests

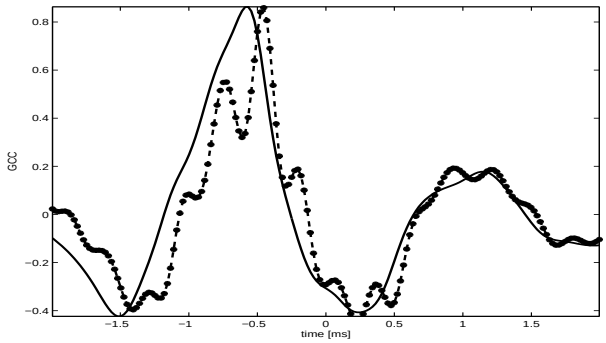
Third, we objectively compare the SSPA and the SHRIR signals by measuring their location with the PHAT-GCC. For cross-pairs (SHRIR-SSPA) at the same position, sounds from SHRIR are perceived more lateralized than sounds of SSPA. This observation is confirmed after the appearance of the cross-correlation function (see Figure 4). The peak of the cross-correlation for SHRIR is positioned left of the one of SSPA for negative angles and right for positive angles. Moreover, we can see the same form of cross-correlation for speech signals and for musical signals. But the SHRIR requires the measurements of HRIRs for all target positions, while the SSPA makes a correct angular interpolation.

The results show a good spatial precision of the SSPA binaural signals. In fact, we observe a dominant peak in the vicinity of the right interaural delay without ambiguity. The cross-correlation function from SSPA are smoother and have fewer parasites than those from SHRIR signals. Thus, the SSPA method seems more accurate and more stable than the SHRIR method. The SSPA method allows to accurately spatialize monophonic sources (voice, instrument). However, we note that the speech signals show a peak broader than the signals from instruments (see Figure 4).

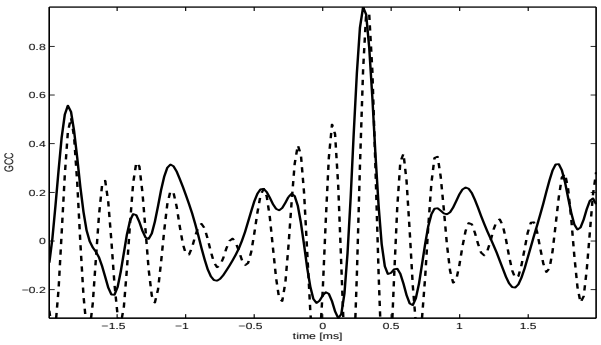
5.3. MSPA performance

In this section, the results analysis is based on real binaural signals, registered in the Bonnefont studio with a "phonocasque" (a headphone with microphones encased in earphones). The recorded signals are of three types, namely those derived from diffusion of a real source monophonic source (one speaker at the target location), those from the parametric multi-diffusion MSPA method and those from the classical VBAP method [3].

[1] see URL: `http://www.labri.fr/perso/~mouba/mos.html`
[2] see URL: `http://dept-info.labri.fr/~sm/SMC08/`



(a)



(b)

Figure 4: *Cross-correlation functions as function of time delay, obtained from binaural signals generated with SSPA (plain) and with SHRIR (dotted): speech at -65° (a), trumpet at $+30^\circ$ (b).*

5.3.1. MSPA and VBAP similarities

In the horizontal plane, MSPA and VBAP follow the pair-wise paradigm (two speakers) to produce a virtual source at a target location. VBAP is known and works well in many situations [15]. Its spatialization is controlled solely by the level difference with frequency independent panning coefficients. Theoretically, VBAP is suitable for frequencies below 700 Hz, which could be sufficient since the ITD, which is an important indication of the location, dominates up to about 1.5 kHz. The panning coefficients of VBAP are fixed for each azimuth and whatever room diffusion. The MSPA panning coefficients are also static regardless of the environment, but they are complex. The MSPA is theoretically defined over the entire audible frequency band. Thus, VBAP is the best candidate who is close to our expectations, and by which we can assess objectively our proposed system MSPA.

Due to the pair-wise paradigm, the comparison of spatialization coefficients for a pair of speakers is sufficient. In the next sections, we used the pair located at $(-30^\circ, +30^\circ)$ for the calculation of panning coefficients for any azimuth between the speakers with both techniques (MSPA and VBAP).

5.3.2. Subjective tests

We conducted listening and objective tests on real sources and on virtual sources from MSPA and VBAP. In all cases, the real source provided a better audio quality and its location is unambiguous. Thus, the real source is considered as the reference (the best we

could expect). Listening tests reveal that the spatial precision of MSPA and VBAP are similar and show no ambiguity, the source is considered properly at the left or right of a previous one for a resolution of 5°. However, the spectral content is different. The virtual source after MSPA has more high frequency contents (brightness), while VBAP sources sound louder. A possible reason is the validity of the VBAP assumption up to 700 Hz. Until 1500 Hz, the ITD cue dominates, which could be sufficient to give a spatial illusion. This observation shows that MSPA should better control the spatialization of broadband components. In fact, Figure 5 shows that the optimal panning coefficients are frequency dependant and not constant over the frequency band.

We also create dynamic sources for a octophonic system with VBAP and MSPA. To have a constant amplitude for any location, we normalize the panning coefficients such that their square sum to 1. The sounds from VBAP seem to have a more constant sound intensity when the source is moving (around the listener).

However, for the two approaches, some acceleration between speakers were reported, with a bias towards the loudspeaker closest to the target location. This effect could be moderate with an increasing number of speakers and a reduction of the angle between each pair of speakers. One advantage of a pair method is that the spatialization error is bounded between the two speakers.

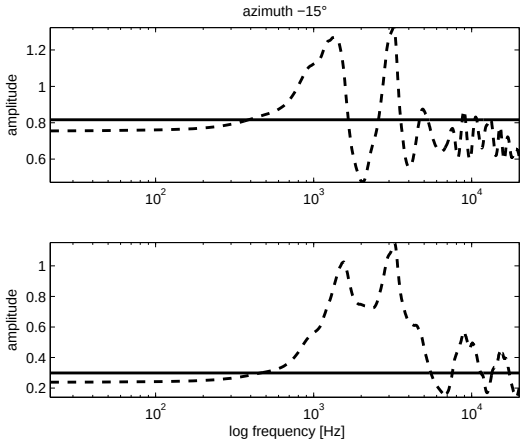


Figure 5: *Amplitude of the panning coefficients from VBAP (plain) and MSPA (dotted), for the left (top) and right (bottom) channels of the panning pair for* -15°.

5.3.3. Objective tests

First, we note that the spatialization coefficients of MSPA and VBAP approaches are very similar up to 700Hz, then they differ considerably (see figures 5 and 6). Indeed, MSPA coefficients are complex numbers and the imaginary part can contribute significantly. In $[0, 700]$ Hz band, the coefficients are nearly real. Over the full band, The Panning Level Difference (PLD) is defined as the ratio of the left by the right panning coefficient. The PLDs difference between the two techniques do not exceed 3 dB in the frequency band $[0, 700]$ for azimuths in range $[-80^\circ, +80^\circ]$ [1].

Moreover, from the binaural recording, it is possible to obtain objective measures of the accuracy on localization. We use binaural signals from the diffusion of a white noise from a real source and virtual sources by MSPA and VBAP at different locations. The

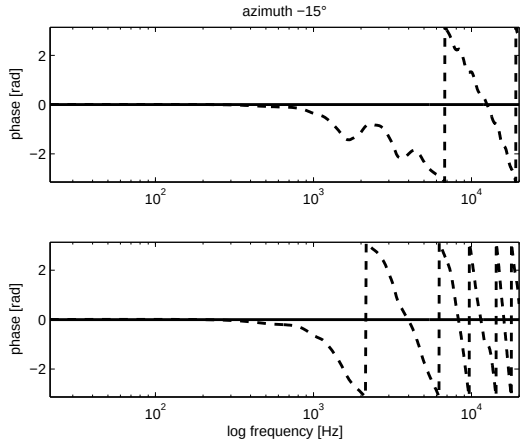


Figure 6: *Phase of the panning coefficients from VBAP (plain) and MSPA (dotted), for the left (top) and right (bottom) channels of the panning pair for* -15°.

white noise is chosen because of its large and constant spectrum, here we used a sampling rate of 44.1 kHz.

The ITD and the deduced azimuth are good accuracy criterions for comparison purposes in source localization [16]. Thus, we believe that the generalized PHAT-GCC is a neutral method for the evaluation of MSPA and VBAP. The final location corresponds to the maximum of the derived cross-correlation function. Estimates of the ITD and the corresponding azimuth can be negative or positive depending on whether the source is positioned towards the left or right ear. The Figure 7 shows the cross-functions obtained from the broadcasting of real sources and virtual sources from MSPA.

We notice that the PHAT-GCC functions of real sources are more accurate and localized at the right position, while the functions of virtual methods present a second significant unwanted peak. However, the dominant peak is still a good estimator of the expected location. The parasites peaks could be explained by complex interactions resulting from the use of two speakers (*e.g.*cross-channels. Table 1 summarizes the results of localization for the azimuths $-30^\circ, -15^\circ, +15^\circ, +30^\circ$. Indeed, the location deduced from the diffusion of mono source is the best we could expect in the acoustics of the room. The results of the three approaches confirm the superiority of the location of the real source. For negative angles, VBAP and MSPA suffer a bias towards the speaker on the left, and for positive angles, the bias moves toward the right speaker. Moreover, we note an localization gain of approximately 2° for MSPA compared to VBAP (see table 1). These findings seem to confirm that MSPA reinforces the correlation between the ITD and ILD in the binaural signals, enabling them to better approximate the ones from natural perception.

5.4. Localization results

To verify the precision of the source localization methods, we spatialized several noise sources at different azimuths in the horizontal plane, between -80° and $+80^\circ$, and we localized them using the conjoint method and the PHAT-GCC method. In these examples, the binaural signals include one single source, such that the PHAT-GCC is not disturb by concurrent sources. We compare the two methods in anechoic environment and in noisy reverberant room.

azimuth θ	real source $\hat{\theta}$	MSPA $\hat{\theta}$	MSPA $\hat{\text{ITD}}$ ms	VBAP $\hat{\theta}$	VBAP $\hat{\text{ITD}}$ ms
-30°	-25°	-27°	-0.22	-27°	-0.24
-15°	-12°	-20°	-0.18	-22°	-0.20
0°	$+1^\circ$	$+1^\circ$	0.01	$+2^\circ$	$+0.02$
$+15^\circ$	$+13^\circ$	$+19^\circ$	-0.18	$+22^\circ$	$+0.20$
$+30^\circ$	$+27^\circ$	$+27^\circ$	-0.24	$+27^\circ$	$+0.24$

Table 1: *Azimuth estimations with PHAT-GCC from binaural signals issued from the diffusion of a white noise as real source or as virtual source generated by MSPA and VBAP using the speakers pair (-30°, $+30^\circ$). Recording with the "phonocasque" in the Bonnefont studio.*

All test files for localization are available here [3] .

5.4.1. In anechoic room

The results of localization for the conjoint method and the PHAT-GCC method are summarized in Figure 8, where the expected azimuths are plotted against the estimated azimuths. Both methods become less accurate as the source gets closer to lateral positions. This phenomenon is also observed in real listening tests, where side sounds were more difficult to locate in absolute.

As we can see, both approaches are almost perfect in the range $[-45, 45]^\circ$ with a maximum error about 3° (see Figure 8). Beyond $|45^\circ|$ both methods become gradually unstable. We remark that the absolute error is less than 5° in the range $[-65, +65]^\circ$. The conjoint method has a error lower than its protagonist. This performance is qualitatively acceptable compared to the human auditory system, which detects differences of 1° [17]. In practice, the source is not a point (but is expanding its activity around a set of points), the size of the speaker and the source's intensity may influence this minimal detectable angle.

Similar tests were conducted on sources with different spectral content, including speech and music. Due to their low frequencies spectrum, the localization results were slightly better than in case of noise signals. We used a noise signal for the automatic detection of the speakers configuration in the RetroSpat Music software [1].

5.4.2. In noisy and reverberant classroom

In this section, we present the results obtained by theconjoint method and the PHAT-GCC method in a reverberant environment. The binaural signals are generated by the convolution of mono sources with BRIRs measured in a reverberant classroom of size $5\text{m} \times 9\text{m} \times 3.5\text{m}$ [18]. The BRIRs have been measured in the horizontal plane from a maximum length sequence, which is a pseudo-random binary sequence. They have a length of 32767 samples and contains a combination of direct sound, first echoes and late reverberation. The reverberation time of room (T_{60}) is between 580 and 619 ms with algorithms of Brown and Schroeder [19].

We study the sources at positions $0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 90^\circ$. The results of localization for the conjoint method and the PHAT-GCC method are summarized in Figure 9. As well as in the anechoic case, we remark that the performance of both techniques decreases as the source moves towards lateral sides.

Although, the PHAT-GCC method is more precise in reverberant environments. Indeed, for the conjoint method, starting from 60°, the error has already reached 10°. For such adverse reverberant conditions, the method is doing rather well. One has the impression that the joint method underestimates the localization; thus a improvement would be to introduce a bias growing with the lateral position.

[3] see URL: `http://www.labri.fr/~mouba/DAFX09.html`

6. CONCLUSION

In this paper, we evaluated the performance of the proposed SSPA and MSPA spatialization methods, respectively for the binaural and the multi-diffusion context. Both methods are based on parametric models of ILD and ITD cues. We also show that our adapted conjoint localization method has comparable precision to the PHAT-GCC (generalized cross-correlation with phase transform) localization method. Our method uses conjointly the localization estimations based on ILD and on ITD at high frequencies. The conjoint approach has the advantage of localizing each frequency component separately. It opens views to locate multiple sources in the same time window, therefore a possible source separation under reverberant conditions. Future works will address the enhancement of the localization algorithm and the problem of multiple sources tracking.

7. REFERENCES

[1] J. Mouba, S. Marchand, B. Masencal, and J-M. Rivet, "Retrospat: a perception-based system for semi-automatic diffusion of acousmatic music," in *Proceedings of the Sound and Music Computing (SMC)*, Berlin, Germany, Sept. 31- Aug. 3, 2008, pp. 33–40.

[2] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Proc. Digital Audio Effects (DAFx-06)*, Montréal, Canada, Sept. 18- 20, 2006, pp. 233–238.

[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.

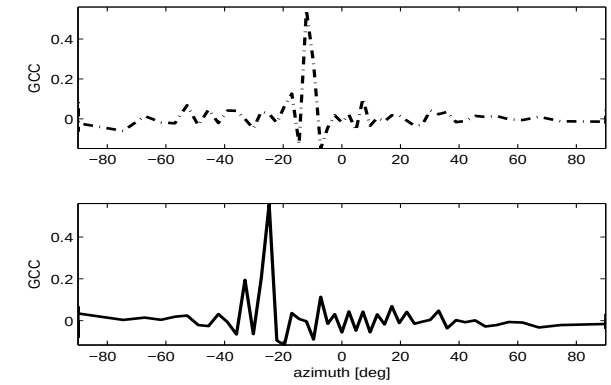
[4] C. H. Knapp and G. C. Carter, "The generalized correlation method for the estimation of time delay," *IEEE Trans. on Sig. Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[5] Francis Rumsey, *Spatial Audio*, Focal Press, Oxford, United Kingdom, first edition, 2001, Reprinted 2003, 2005.

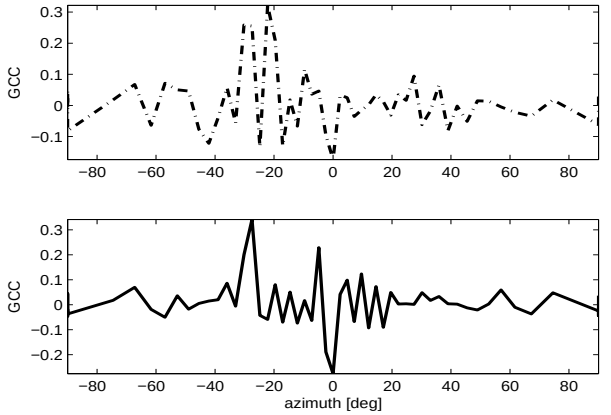
[6] J. Blauert, *Spatial Hearing*, MIT Press, Cambridge, Massachusetts, revised edition, 1997, Translation by J. S. Allen.

[7] J. W. Strutt (Rayleigh), "On our Perception of Sound Direction," *Philosophical Magazine*, vol. 13, pp. 214–302, 1907.

[8] H. Viste, *Binaural Localization and Separation Techniques*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 2004.



(a)



(b)

Figure 7: *Cross-correlation functions as function of azimuth, obtained from binaural signals generated with real source (a) and with MSPA diffusion in a reverberant environment (Bonnefont studio), at -15° (dashed lines) and -30° (solid lines).*

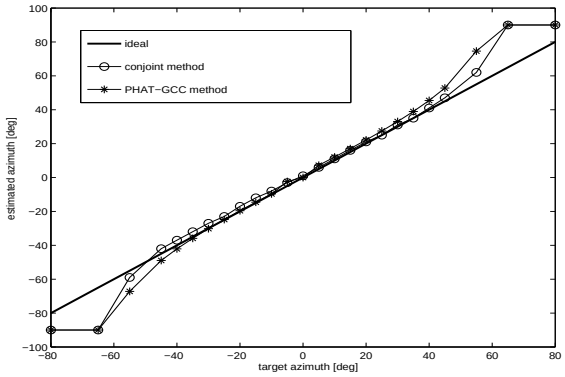


Figure 8: *Source localization, in a anechoic room, of a white noise at different azimuths with different methods: ideal (plain), conjoint method (round), PHAT-GCC method (asterisk).*

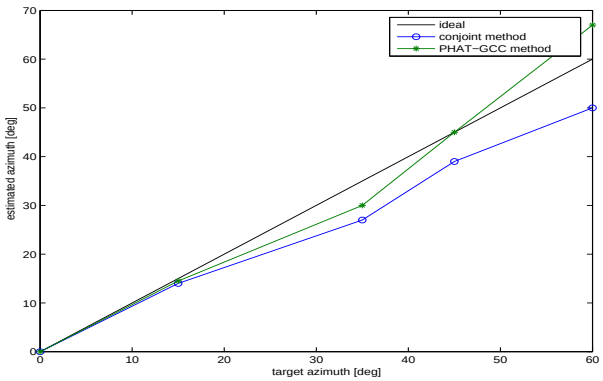


Figure 9: *Source localization, in a reverberant classroom, of a white noise at different azimuths with different methods: ideal (plain), conjoint method (round), PHAT-GCC method (asterisk).*

[9] 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, 2001, pp. 99–102.

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

[11] George 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.

[12] C. Tournery and C. Faller, "Improved Time Delay Analysis/Synthesis for Parametric Stereo Audio Coding," *Journal of the Audio Engineering Society*, vol. 29, no. 5, pp. 490–498, 2006.

[13] John M. Chowning, "The Simulation of Moving Sound Sources," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 2–6, 1971.

[14] Arthur N. Popper and Richard R. Fay, *Sound Source Localization*, Springer, New York, 2005.

[15] V. Pulkki, "Generic panning tools for MAX/MSP," in *Proceedings of the International Computer Music Conference*, Berlin, Germany, August 2000, pp. 304–307.

[16] E. Bates, G. Kearney, F. Boland, and Dermot Furlog, "Localization accuracy of advanced spatialization techniques in small concert halls," in *153rd meeting of the Acoustical Society of America*, June 2007.

[17] William M. Hartmann, "How We Localize Sound," Available at http://www.aip.org/pt/nov99/locsound.html, accessed April 05, 2009.

[18] B.G. Shinn-Cunningham, N. Kopco, and T Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117.

[19] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. of Am.*, vol. 37, pp. 409–412, 1965.