SOUND SOURCE LOCALIZATION FROM INTERAURAL CUES: ESTIMATION OF THE AZIMUTH AND EFFECT OF THE ELEVATION

Éric Méaux[1] **Sylvain Marchand**[1]

[1] L3i, University of La Rochelle, France

eric.meaux1@univ-lr.fr, sylvain.marchand@univ-lr.fr

ABSTRACT

Localizing a sound source in various contexts represents a complex task, and numerous methods have been proposed for this purpose.

Human beings use acoustic cues (like Interaural Time Differences and Interaural Level Differences) to handle this task, which is also the case for the localization method we propose, after Viste. More precisely, we model the cues as functions of the azimuth, then we inverse the models to estimate this azimuth from the measured cues obtained from binaural signals.

This theoretical background is used by our STAR method (Synthetic Transaural Audio Rendering), a spatialization technique based on the reconstitution of interaural cues at the ears of the listener, from a selected pair of loudspeakers using a transaural technique with synthetic Head-Related Transfer Functions.

However all these localization methods focus on the azimuth, in the horizontal plane, and do not take into account the elevation of the sound source, thus raising the following problematic: Does elevation impact the estimation of the azimuth?

Indeed, it is essential to check if the azimuth is still efficiently estimated regardless of the elevation of the sound source. For example, complex spatialization systems, such as the ones relying on a dome of loudspeakers, produce sounds at different elevations, above and below the ears of the listener.

The proposed work studies to which extent elevation impacts the precision of the estimation in azimuth, with in mind a generalization to 3D of both localization and spatialization existing methods.

The paper starts by describing the method previously implemented, then emphasizes the experimental tests performed with the KEMAR manikin placed at the center of a dome of loudspeakers. Finally, the different results will be presented and commented, and we will discuss how the existing methods can be extended to 3D.

1. INTRODUCTION

The purpose of localizing an audio source is to ascertain where a sound is actually produced. This problematic is well known and several methods have been proposed [1,2].

Among these methods, a perceptive approach for sound localization has been introduced [3, 4]. The focus is not on physical factors but rather on human ones. Furthermore a head and torso system is necessary to record the sound. Indeed this model is based on acoustic interaural cues, used by the human auditory system for sound localization [5,6].

The aim is to use these interaural cues and then to process them in a human-like fashion. This must work with various head and torso manikins and in realistic configurations. In this paper, the original method is tested in different conditions, using a Head-Related Impulse Response (HRIR) database, a Binaural Room Impulse Response (BRIR) database, and dedicated sound recordings using a KEMAR manikin.

In Section 2 the method is introduced, while Section 3 presents the experiments conducted to test the method resistance. Different conditions have been taken into account: anechoic conditions with a HRIR database [7], reverberant room conditions with a BRIR database [8], as well as dedicated recordings in our SCRIME studio. These tests show than the method localizes the sound source in azimuth independently of the elevation, thus paving the way for a 3D implementation (as future work).

2. METHOD

The interaural cues, namely ITD (Interaural Time Difference) and ILD (Interaural Level Difference), is a concept now standard in the acoustical community. Based on them, methods for sound localization and spatialization were developed [3, 4, 9].

2.1 ILD and ITD Models

The ITD (Interaural Level Difference) represents the travel time difference of a sound between the two ears while the ILD (Interaural Level Difference) represents the level difference. ITD and ILD are the main acoustic cues used by human beings to localize a sound. It turns out that ILDs are more efficient at high frequencies while ITDs are prominent at low frequencies [10]. Physically, high frequencies

are more sensitive to frequency-selective amplitude attenuation, even if the associated signal is phase ambiguous. On the other hand, low frequencies have no ambiguity but are less sensitive [5].

The localization method proposed for the present study relies on ILD and ITD models of Eqns. (1) and (2), proposed by Mouba et al. [3] based on Viste [11]:

$$\mathrm{ILD}_{\mathrm{model}}(\theta, f) = \alpha(f)\sin(\theta) \quad (1)$$

$$\mathrm{ITD}_{\mathrm{model}}(\theta, f) = \beta(f) r \sin(\theta)/c \quad (2)$$

where α and β are frequency-dependent scaling factors (see [11] and Fig. 1), that encapsulate the head / ears morphology. In our experiments, we use the mean of individual scaling factors over the 45 subjects of the CIPIC database. For each subject, we measure the interaural cues from the Head-Related Transfer Functions (HRTFs, which are the spectral version of the HRIRs) and derive the individual scaling factors that best match the model – in the least-square sense – for all azimuths.

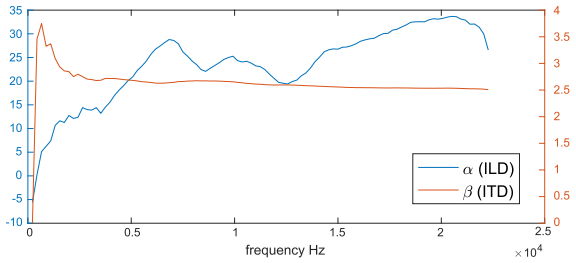

Figure 1. α and β scaling factors.

Given the spectra of the left (L) and right (R) channels, we can estimate the ILD and ITD with:

$$\mathrm{ILD}(f) = 20\log_{10}(|L(f)/R(f)|) \quad (3)$$

$$\mathrm{ITD}_p(f) = \frac{1}{2\pi f}(\angle(L(f)/R(f)) + 2\pi p) \quad (4)$$

The coefficient p outlooks that the phase is determined up to a modulo 2π factor.

2.2 Localization Process

The purpose of the models previously introduced is to simply recreate acoustic cues (relying on sound source azimuth, frequency, and scaling factors). The localization model is based on the assumption that if a sound is recorded using a head and torso system, such models allow to estimate the azimuth θ (the only unknown variable). An overview of this process is described on Fig. 2. For the left and right input signals sampled at 44.1kHz, frames of $N = 2048$ samples are considered.

Part a : The ILD and ILD are calculated according to Eqns. (3) and (4).

Part b : One azimuth estimate is found from the ILD. Indeed the scaling factor α is known as well as the ILD. Then by inversion of Eqn. (1) we get Eqn. (5) and deduce θ_{ILD} from the ILD:

$$\theta_{\mathrm{ILD}}(f) = \arcsin(\mathrm{ILD}(f)/\alpha) \quad (5)$$

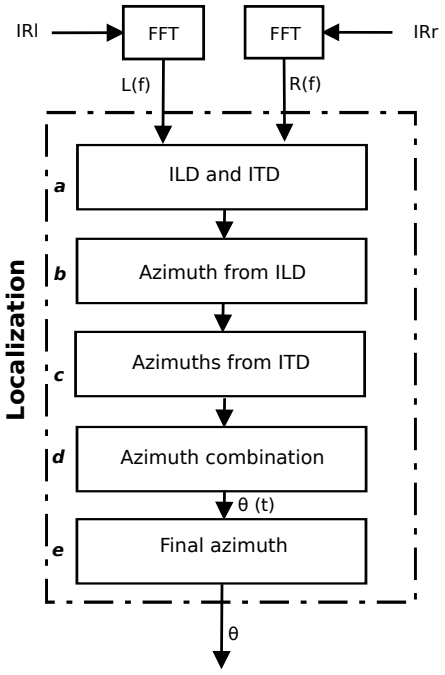

Figure 2. Localization process chain.

Part c : Other azimuth estimates are calculated by inversion of Eqn. (2), but with an ambiguity factor p remaining, using Eqn. (6):

$$\theta_{\mathrm{ITD},p}(f) = \arcsin(\mathrm{ITD}_p(f)c/(\beta r)) \quad (6)$$

Part d : Then, the $\theta_{\mathrm{ITD},p}$ that is the nearest to θ_{ILD} is validated as the final $\theta(f)$ using Eqn. (7):

$$\theta(f) = \theta_{\mathrm{ITD},m}(f) \text{ with } m = \mathrm{argmin}_p|\theta_{\mathrm{ILD}} - \theta_{\mathrm{ITD}_p}| \quad (7)$$

Part e : In theory, a single source should give the same azimuth for all frequencies. In practice, the presence of noise and reverberation spreads the azimuth energy. The estimated azimuth θ is chosen as the location of the peak of the histogram of the distribution of this energy, as shown on Fig. 3. Since it turns out that the localization method

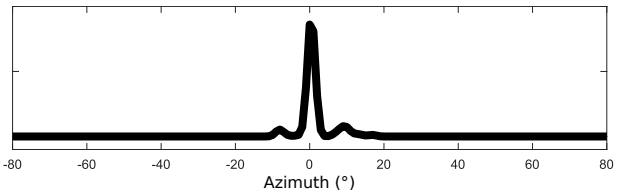

Figure 3. Azimuth histogram in ideal conditions (using CIPIC database). The energy is very well concentrated around the sound source position (here 0°).

can sometimes produce extreme azimuths, only the ones within the $\pm$ 80° range are considered for the histogram.

3. METHOD RESISTANCE

The aim of this section is to study the resistance of the previously explained method in different contexts. To conduct

these resistance tests, we first used HRIR [7] and BRIR [8] databases for the purpose of simulation. Both databases use the KEMAR head and torso manikin, also employed in our own study measurements. Indeed a proper set of data was acquired in our studio in order to complete the BRIR database, adding the vertical component. The simulation consists in placing virtually (by convolution) a Gaussian noise sound source at a known position, and then in estimating this position.

3.1 Precision of Azimuth in Anechoic Conditions

The CIPIC database [7] contains HRIRs for 45 subjects (including KEMAR small and large pinnae) at 25 different azimuths and 50 different elevations, acquired in anechoic conditions with Bose Acoustimass loudspeakers.

3.1.1 Estimation of Azimuth in Anechoic Conditions

The first step is to validate the azimuth localization when the sound source is produced on the transaural plane. Fig. 4 shows the estimated azimuth from the method based on five different sound source positions (from -25° to 25°), with HRIRs provided from the KEMAR large pinnae. The estimation error is low, below human precision. Indeed, for a human being, even if localization is a complex subject [5, 12], it is generally accorded that the localization precision is around 5° in azimuth (and 10° in elevation) [13, 14].

Moreover if all the 43 human subjects of CIPIC are considered, the results are quite similar to the ones obtained using the KEMAR large pinnae head and torso manikin (of CIPIC too).

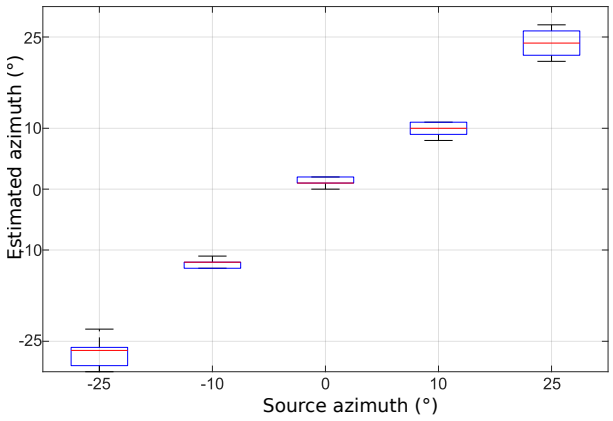

Figure 4. Azimuth localization precision for 5 positions and 43 CIPIC database subjects for all 25 elevations.

3.1.2 Resistance to Elevation in Anechoic Condition

The previous results have shown that the KEMAR head and torso manikin and human subjects provide similar results, thus only the KEMAR large pinnae of the CIPIC database will be further considered.

The computed azimuth according to the elevation (see Fig. 5) shows that the localization precision using our method is very good (considering human precision). Indeed, on this figure, the different dot-lines follow the reference green lines (corresponding to the real audio source position). More interestingly, the elevation does not affect the azimuth estimation precision.

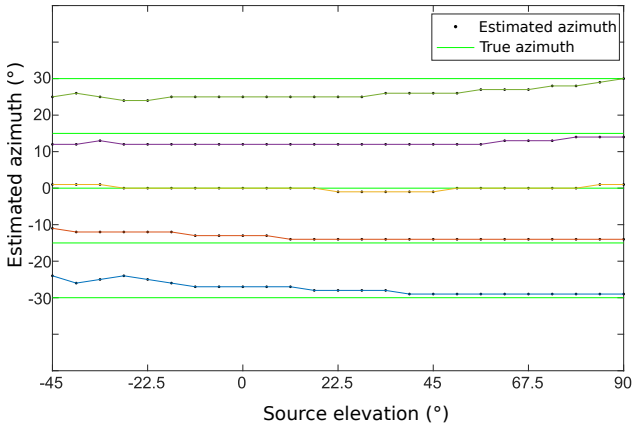

Figure 5. Azimuth localization precision for 5 positions (with KEMAR large pinnae manikin). The estimation follows the real azimuth quite well.

It also interesting to look at the azimuth estimation error for different elevations. Fig. 6 summarizes that. For each elevation, all azimuths are considered. The localization has the expected precision (here $\pm$ 5°). Moreover, the azimuth localization error does not depend on the elevation. The only problem can be found in extreme elevations, which are also problematic for human perception [13, 14] as well as for mathematics themselves (for extreme elevations, azimuth is meaningless since every value leads to the same position).

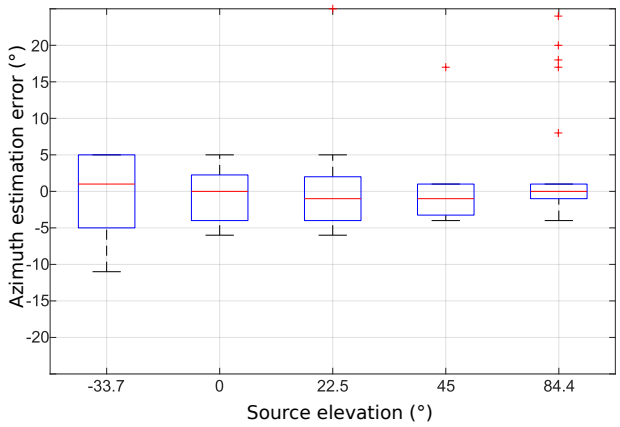

Figure 6. Azimuth localization error for 5 elevations (with KEMAR large pinnae manikin), for all azimuths of the CIPIC database.

3.2 Test in Real Conditions

3.2.1 Estimation of the Azimuth

The previous part discussed the method in anechoic conditions, which are not really realistic. Indeed, anechoic rooms are rather uncommon outside the scientific context. Regarding reverberant rooms, the literature suggests a large set of free databases. For this study, the choice

has been made on the BRIR database of the Audio Lab of the Institute of Communications Engineering (ALICE) [8], University of Rostock. There, 64 Neumann KH 120 A loudspeakers were mounted at ear height on a square truss construction. In our study, only the loudspeaker in front of the manikin will be used. The room have dimensions 5m $\times$ 5.75m, and 3m height. One wall has windows and a wooden door. Walls and ceiling are plastered, only the wall with the door is a drywall. The floor is covered with a thin carpet. To change the room impulse response, some configurations have been varied for the measurement (no absorbers in the room; broadband absorbers at walls and in front windows; broadband absorbers at walls, ceiling and in front of the windows; additional absorbers of pyramid-shaped foam with 7 cm depth). BRIRs have been measured with linear sweeps acquired with a KEMAR large pinnae manikin. The head of the manikin was rotated horizontally above the torso from $\pm 80°$ in $2°$ steps.

The results are comparable to the data acquired in our SCRIME studio, which will be described in the next part.

Fig. 8 illustrates the localized sound using our model, varying the sound source azimuth, for different rooms configurations. In dotted red, the ideal results are shown. The different estimated azimuths follow the ideal in a $\pm 40°$ range. Above, a bias is visible, which could be explained by reflections. Indeed, the reflected source may have a greater energy than the direct source. Fig. 7 provides explanation of that bias: the localization method is mistaken between image source and real source.

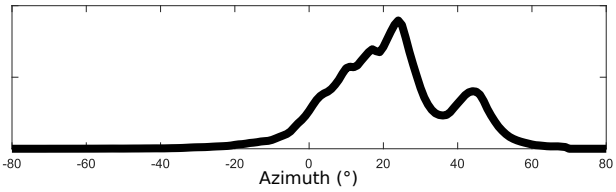

Figure 7. Azimuth histogram in real conditions (using the ALICE BRIR database). The energy is spread and the reflected source (25°) has more energy than the real source (48°).

However, in the $\pm 40°$ range, the localization method works in real conditions and is consistent with the human localization precision [5]. Indeed, even if the localization seems to have a standard deviation more important than in anechoic conditions, the results are in line with those expected.

Fig. 9 shows the azimuth estimation error for all the 41 azimuths ($\pm$ 40 ° range). The results show that the estimation is unbiased (zero-mean error) and that the standard deviation of the error is comparable to human performance, which confirms that the model is valid at elevation $0°$ including real conditions.

That figure is also interesting for comparing the different BRIRs from the different rooms [8] and the SCRIME studio measurement (in first position). The results are quite similar.

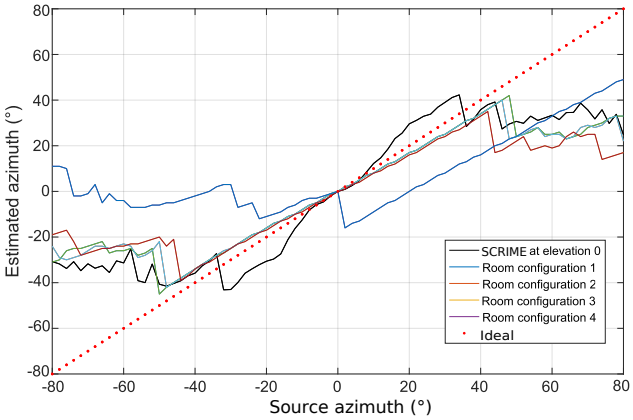

Figure 8. Azimuth estimation for different room configurations. The results follow well the ideal in $\pm$ 40 ° range.

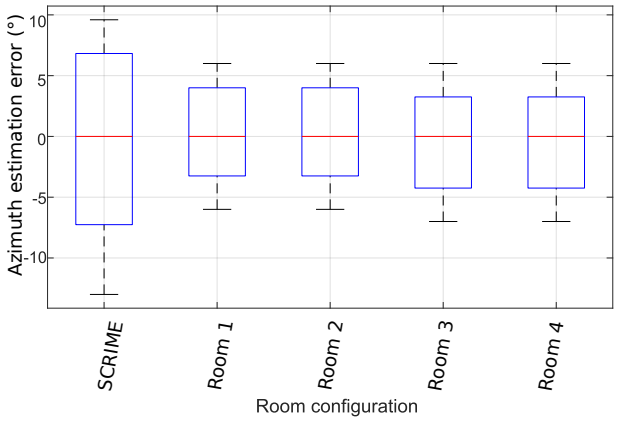

Figure 9. Azimuth estimation error for different rooms, including our SCRIME studio. It should be noted that Room 1 generates a lot of outliers, out of the plot.

3.2.2 BRIR Exploitation in Elevation

The CIPIC HRIR database is very well calibrated and provide a database containing positions both for azimuth and elevation, however the recording conditions are anechoic and thus not realistic. The ALICE BRIR database provides recordings in real conditions, but only for elevation zero. For the present study, we needed some elevation data in non-anechoic room. For this reason, we designed an experimental process, closest as possible to the experimental processes used for the acquisition of the previous databases.

More precisely, the measurements have been performed at the Studio de Création et de Recherche en Informatique et Musiques Expérimentales (SCRIME), University of Bordeaux, France. This studio is used by musicians and has quite good acoustics, even if it is not physically controlled. There, 18 Genelec 8030 loudspeakers are mounted on three loudspeakers rings. The studio has a surface of 40m square. One wall has three windows, two have a wooden door, and some acoustic panels are disposed against them. The floor is covered with a thin carpet. For the source signal we have used a white noise of 6s length, sampled at 44.1kHz, emitted from a unique loudspeaker lo-

cated 2.6m ahead from the KEMAR large pinnae manikin. The manikin was rotated horizontally from $\pm 80°$ in $2°$ steps, and from $\pm 40°$ vertically using some ad-hoc system shown on Fig. 10. This system is mounted on an office chair support, which provides simple rotation and displacement. To simulate the vertical position, the manikin is inclined using strong hinge fixed on the top of the office chair support, giving the advantage to have easily access to a lot of vertical positions. When the manikin is rotated to simulate some elevation it is still possible to rotate it on the azimuth plane. Using the system, and because of its own height, a bias appears when the inclined manikin is rotated in the azimuth plane. Replacing the head on the correct position could generate too much experimental errors. For this reason, a correction is done a posteriori, using manikin size, and known vertical and horizontal angles.

Figure 10. Installation for the $20°$ recording at the SCRIME. Although the manikin is surrounded by loudspeakers, only the one in front of it is used for the measurements.

Even if the SCRIME database is less calibrated, the results are consistent with the ones of the existing databases.

Fig. 11 and 12 show the localization results for the set of source azimuths at five elevations.

The same conclusion can be made as in the study in anechoic conditions (Fig. 6): the elevation does not affect azimuth localization.

4. CONCLUSION AND FUTURE WORK

In this article, a perceptive model for source localization in azimuth was presented and tested.

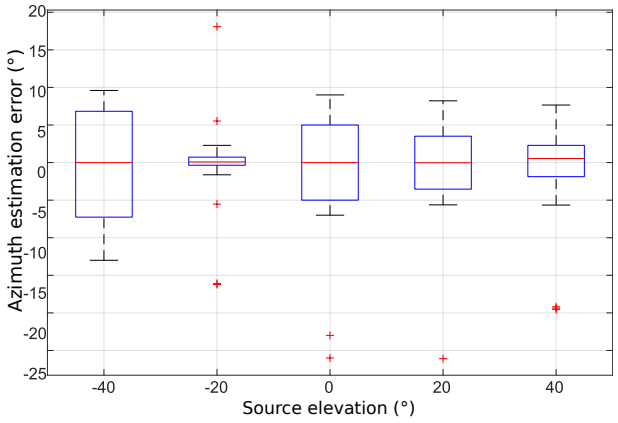

Figure 11. Azimuth error for different elevations with SCRIME recordings.

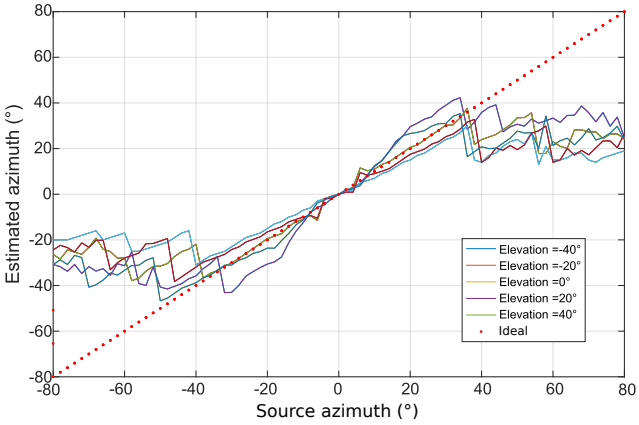

Figure 12. Estimated azimuth for different elevations with SCRIME recordings. The results follow well the ideal in the $\pm$ 40 $^\circ$ range.

The azimuth estimation is comparable to human performance, as shown with our tests using the HRIR database. It is also shown that the method is resistant to both elevation and reverberation, as shown with our tests using the BRIR database and our own experiments at the SCRIME studio.

A problem nevertheless appears with extreme elevations (where the azimuth is meaningless though), as well as in the case where the reverberations have an energy superior to the main source.

This study therefore shows that the method precisely localizes a sound source in azimuth, and this independently of the elevation, and even in realistic conditions (reverberant rooms, noisy environment).

This is the first step to a full 3D method to localize and spatialize sound using acoustic cues.

5. ACKNOWLEDGMENTS

The authors would like to thank the SCRIME (Studio de Création et de Recherche en Informatique et Musiques Expérimentales), University of Bordeaux, France, which supplied the studio and materials for the recording of our database.

6. REFERENCES

[1] C. Beck, G. Garreau, and J. Georgiou, "Sound Source Localization through 8 MEMS Microphones Array Using a Sand-Scorpion-Inspired Spiking Neural Network," *Frontiers in Neuroscience*, vol. 10, p. 479, 2016.

[2] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source Counting in Real-Time Sound Source Localization Using a Circular Microphone Array," in *Proc. of the 7th IEEE Sensor Array and Multichannel Signal Processing*, (Hoboken, NJ, United States), pp. 521–524, 2012.

[3] J. Mouba, S. Marchand, B. Mansencal, and J.-M. Rivet, "RetroSpat: a Perception-Based System for Semi-Automatic Diffusion of Acousmatic Music," in *Proc. of the Sound and Music Computing (SMC) Conference*, (Berlin, Germany), p. 33–40, July/August 2008.

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

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

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

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

[8] V. Erbes, M. Geier, S. Weinzierl, and S. Spors, "Database of Single-Channel and Binaural Room Impulse Responses of a 64-channel Loudspeaker Array," in *Proc. of the 138th Convention of the Audio Engineering Society (AES)*, (Warsaw, Poland), May 2015.

[9] Éric Méaux and S. Marchand, "Synthetic Transaural Audio Rendering (STAR): a Perceptive Approach for Sound Spatialization," in *Proc. of the International Conference on Digital Audio Effects (DAFx)*, (Birmingham, United Kingdom), September 2019.

[10] L. Rayleigh, "On the Perception of the Direction of Sound," *Proc. of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 83, no. 559, pp. 61–64, 1909.

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

[12] J. C. Middlebrooks and D. M. Green, "Sound Localization by Human Listeners," *Annual Review of Psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[13] M. Risoud, J.-N. Hanson, F. Gauvrit, C. Renard, P.-E. Lemesre, N.-X. Bonne, and C. Vincent, "Sound Source Localization," in *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 135, pp. 259–264, 2018.

[14] T. Letowski and S. Letowski, "Localization Error: Accuracy and Precision of Auditory Localization," *Advances in sound localization*, pp. 55–78, 2011.