Magnitude-Squared Coherence (MSC) – the MSC $\gamma[k]$ measures the linear correlation between the left and right signals as a function of the frequency [Vesa, 2007, 2009]. Since reverberations are known to provide distance information, but to highly degrade this binaural coherence, the MSC is expected to be reliable as a distance cue. It is defined along

$$\gamma[k]=\frac{\mid G_{lr}[k]\mid^2}{G_{ll}[k]G_{rr}[k]},\ \text{with} \tag{3.33}$$

$G_{lr}[k]=\langle L^*[k]R[k]\rangle$, $G_{ll}[k]=\langle|L[k]|^2\rangle$, $G_{rr}[k]=\langle|R[k]|^2\rangle$, and $\langle.\rangle$ a first order leaky integrator verifying $\langle Q[t]\rangle=\beta\langle Q[t-1]\rangle+(1-\beta)Q[t]$ (t denotes the frame index and β is set to 0.5 in all the following). Note that $\gamma[k]$ definition is very close to the square of the PHAT whitening fonction $W_M[k]$ (see Equation 3.46), with a time filtering step added to its definition.

DRR based on Spatial Correlation Model (DRR-SCM) –DRR-SCM has been proposed in [Hioka et al., 2010, 2011] for array processing approaches with multiple microphones. The method is *********applied here in a binaural context and exploited to provide frequency- dependent DRRs, which can provide more details since the wall sound absorptions and reflections are frequency-dependent*********. First, the transfer function between the sound source and the left or right signals is decomposed into frequency-dependent direct $H_D[k]$ and reverberant $H_R[k]$ components. Thus, one can write:

$$L[k] \quad = \left(H^l_D[k]+H^l_R[k]\right)S[k], \tag{3.34}$$

$$R[k] \quad = \left(H^r_D[k]+H^r_R[k]\right)S[k], \tag{3.35}$$

with $S[k]$ the FFT of the source signal. Consequently, the cross correlation $R_{lr}[k]$ between the left and right channels comes as

$$\begin{aligned} R_{lr}[k] \triangleq \mathbb{E}\left[L[k]R^*[k]\right] = \mathbb{E}\left[|S[k]|^2\{H^l_D[k]H^{r*}_D[k]+\right.\\ \left.H^l_R[k]H^{r*}_R[k]+H^l_D[k]H^{r*}_R[k]+H^l_R[k]H^{r*}_D[k]\}\right], \end{aligned} \tag{3.36}$$

with $\mathbb{E}[.]$ the mathematical expectation. The direct and reverberant components are supposed to be made of plane waves (this also involves that the head HRTF is neglected), so that $|H^l_D[k]|=|H^r_D[k]|=|H_D[k]|$ and $|H^l_R[k]|=|H^r_R[k]|=|H_R[k]|$. The reverberant component is also hypothesized as being diffuse, and the cross-correlation between the direct and reverberant components is assumed to be sufficiently small. Then, the spatial correlation matrix $\mathbf{R}[k]$ of the binaural signals comes as [Hioka et al., 2010]:

$$\begin{aligned} &\mathbf{R}[k] \simeq P_D[k]\begin{pmatrix}1 & d_{lr}\\ d_{rl} & 1\end{pmatrix}+P_R[k]\begin{pmatrix}1 & r_{lr}\\ r_{rl} & 1\end{pmatrix},\\ &\text{with } d_{lr}=1/d_{rl}=\exp\left(j2\pi\frac{kF_s}{Mc}\text{ITD}\right),\\ &\text{and } r_{lr}=r_{rl}=\text{sinc}(2\pi\frac{kF_sa}{Mc}), \end{aligned} \tag{3.37}$$

where M is the number of points of the FFTs $L[k]$ and $R[k]$, a the distance between the two ears, $P_D[k]=\mathbb{E}\left[|S_M[k]|^2|H_D[k]|^2\right]$ represents the (unkown) direct-path power spectral density (PSD) function, and $P_R[k]=\mathbb{E}\left[|S_M[k]|^2|H_R[k]|^2\right]$ the (unknown) reverberant one. Note that

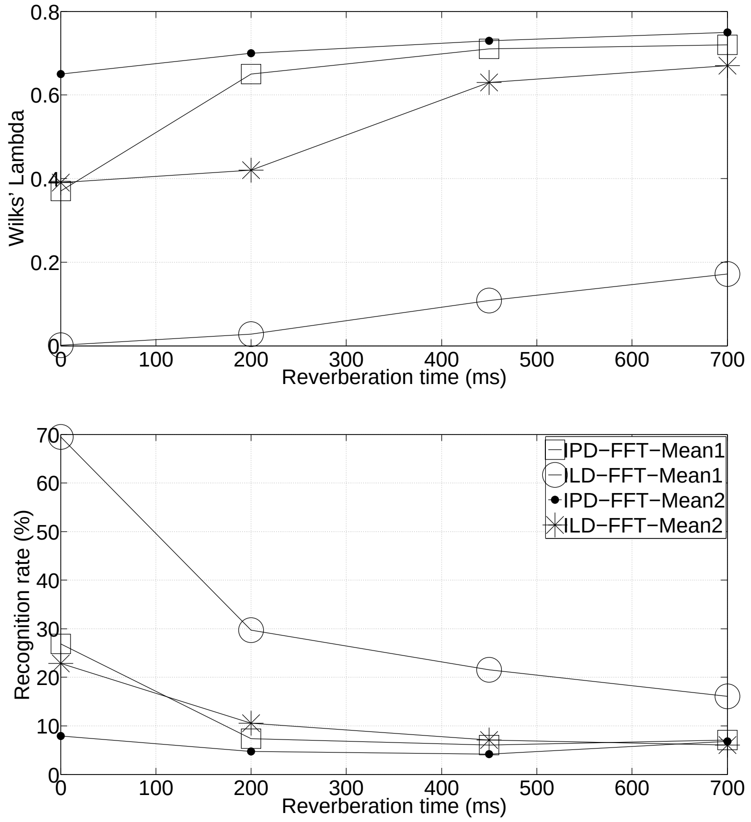


Figure 3.12: Wilks Lambda measures (up) and recognition rates (down) for multiple azimuth FFT-related cues computation techniques as a function of reverberation times. The notations used in the legend are as specified in § 3.4.1.

discriminating than ILDs. Thus, in the following, only the "FFT-Mean1" approach is used and compared with the other cue extraction techniques.

Overall azimuth cues – this paragraph shows the results of the study made with the ensemble of the cues reviewed in § 3.4.1. Wilks' Lambda and recognition rates are shown in Figure 3.13. In this figure, the cues are as denoted in § 3.4.1: ITD_{CC}, ITD_{GCC}, ILD, IPD_{FFT}, ILD_{FFT} (computed in the FFT-Mean1 approach, and ITD_{coch} and ILD_{coch} (computed with no modification on the outputs of gammatone filters). Note that FFT and cochlea-based cues are computed on 30 frequency bands. Generally, and as expected, the recognition rate increases as Wilks' Lambda decreases, which provides logical conclusions despite the simplicity of the LDA-based classifier. In summary, the cues computed using the original temporal signals with no frequency dependent information exploitation present the weakest discriminatory abilities. At the same time, ITDs and ILDs extracted after cochlear filtering are better than IPDs and ILDs extracted based on the FFTs, according to the current implementations.

Distance cues – in the distance case, and having distances between $1m$ and $2.8m$ with $45cm$ steps, 5 groups are obtained. The evaluated cues are Vesa's mean-squared coherence [Vesa, 2007], [Vesa, 2009] (MSC), Smaragdis's relative phase and magnitude [Smaragdis and Boufounos, 2007] (RP and RM), Lu's DRR-EQ [Lu and Cooke, 2010], *******and

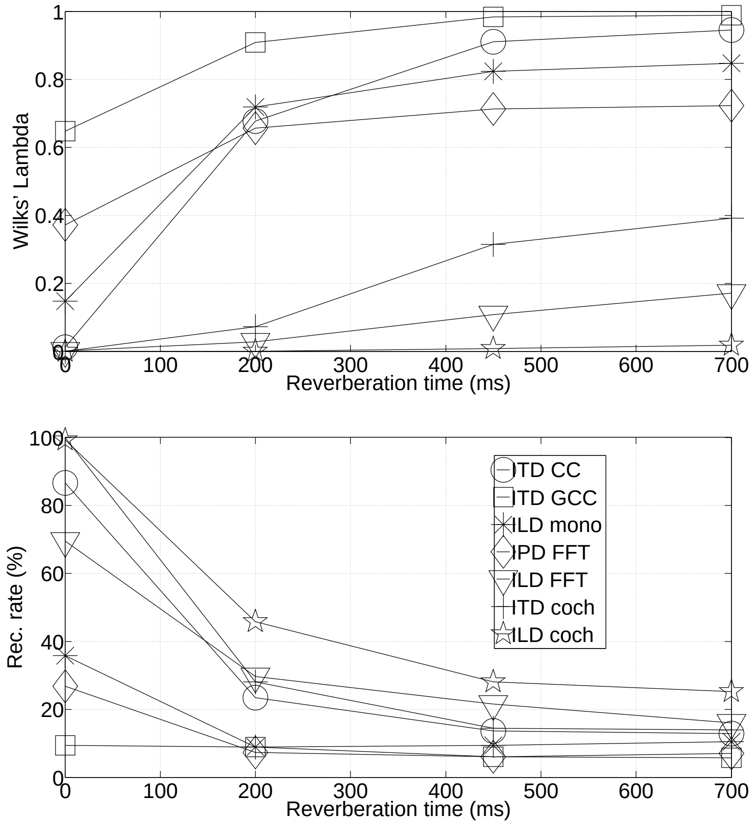


Figure 3.13: Wilks' Lambda measures and LDA-based recognition rates for multiple azimuth-related cue computation techniques for multiple reverberation times.

Hioka's DRR-SCM cue***** [Hioka et al., 2010, 2011], all used in 30-d vectors, being computed on 30 frequency bands. Indeed, the first two provide as many cues as used FFT points, which leads to very large cue dimensions. The current implementation takes 30-D vectors after averaging the obtained cues on 30 consecutive equal frequency bands covering the available frequency range. At the same time, Lu's DRR-EQ computation provides a single DRR value per frame, the dimension has been increased by computing DRRs on 30 similar frequency bands also. It has been reported that some distance cues are azimuth-dependent, leading to different distance estimation performances for different azimuths. Results obtained after cue computation on speech signals for both azimuths of 0° and 45° separately are reported in Figure 3.14. A first observation of the plotted results shows that all the cues are the least effective for a 0ms reverberation time. Wilks' Lambda measures decrease when passing from anechoic to echoic conditions. This shows the need of reflections for distance estimation. But the performances vary according to the RT60. Indeed, for most of the cues and from a certain RT60 value, a higher RT60 does not imply better distance discrimination ability, but the contrary. In all cases, *******the DRR-SCM cue** ([Hioka et al., 2010, 2011])******* is shown to be the most effective and robust to RT60 and source azimuth changes. This shows that the spatial correlation model, despite using an assumption of diffuseness of the environment, only reached in ideal situations, is a useful model providing reliable information for the method, as well as the decomposition of the impulse response linking the emitted and received sounds into direct and reverberant components. This cue will thus be adopted for a system outputting the

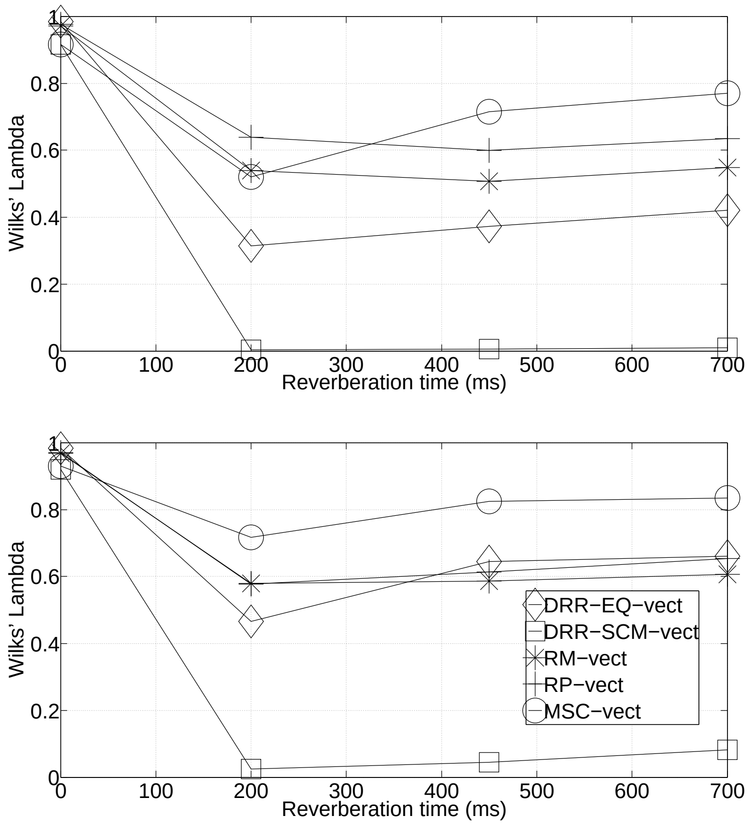


Figure 3.14: Wilks' Lambda measures for multiple distance-related binaural cues in function of the reverberation time and two azimuths: 0° (up) and 45° (down).

value of the source distance, to be seen in the next chapter.

3.4.5 Discussion

This section summarized and evaluated different techniques used in the literature for computing azimuth and distance binaural cues. A comparison of the position discriminatory powers of these cues showed the differences in qualities that they have. Such a study is much needed as a means for better understanding the state-of-the-art and deciding about techniques to use in such a system. In conclusion, the gammatone filtering resulting cues seem to be the most appropriate for azimuth estimation. Distance is still not widely addressed and binaural distance cues show a sensitivity to the source's azimuth, *********but Hioka's distance cue DRR-SCM is robust and better discriminating distances than others*********. These conclusions will be taken benefit of in the established azimuth and distance estimation systems, that will be presented in the following section.

3.5 Conclusion

In this chapter, acoustic cues for sound source localization have been addressed. The human auditory system functioning regarding this task has been overviewed. The task can be divided into azimuth, elevation, and distance estimation which are done through different cues, affected by both the human listener's body, and the room acoustics. Then, a literature overview of the methods implemented to localize sources in the binaural context has been made. It showed that the computation of localization cues can be done through multiple techniques and the question to ask was which technique provides the best cues among them all. Thus, a systematic study implementing all the techniques on the same data was done. It showed that:

- for azimuth estimation, interaural cues computed using gammatone filterbanks,
- for distance estimation, *********the DRR computation technique dubbed DRR-SCM*********, provided the best cues for discriminating different source-receiver positions. These conclusions will be taken benefit of in the next chapter where a sound source localization system will be proposed.

Chapter 4

Binaural sound source localization: proposed approach

Works conducted in Chapter 3 concluded that using gammatone filterbanks before interaural cue extraction provided better azimuth discrimination than other azimuth cue extraction techniques. Also, *********the DRR-SCM cue, presented in** [Hioka et al., 2010, 2011] **and shown in § 3.3.3*********, led to better distance discrimination capabilities than other distance cue computation techniques. We have now a solid indicator on which cue extraction techniques are to adopt among the analyzed techniques in Chapter 3, in a system estimating the sound source position. These conclusions are taken into consideration in the current chapter as the azimuth, distance and elevation estimation tasks are addressed, with respective approaches presented, optimized and evaluated with simulation and real recorded data.

4.1 Azimuth and distance estimation, proposed system

As shown in § 3.3, the state-of-the-art concerning sound source localization provides multiple studies attempting to localize sound sources, mostly in terms of source-receiver azimuth and distance estimation. But we showed that these techniques disagree on multiple components and parameters that affect and steer their performances. In this section, the established azimuth and distance estimation approaches are presented. They follow the general structure shown in § 3.3.1 and Figure 3.6. The signals of the two ears are exploited to obtain cues that enter in pattern recognition approaches outputting the estimated coordinate (azimuth or distance). We try to be better placed than previous localization works by taking benefit of the study made in the previous chapter, to adopt our cue computation techniques, and pursuing the system conception in the same logic to optimize some of its parameters. The estimation approaches are first presented, some of their parameters are optimized later.

4.1.1 Azimuth estimation approach

For the azimuth estimation, and relying on the conclusions of the study conducted in the previous section, the computed cues are interaural time and level differences based on gammatone filterbank outputs simulating the analysis happening in the cochlea. These cues are input to a neural network trained to output the source azimuth. The state-of-the-art

4.1.2 Distance estimation approach

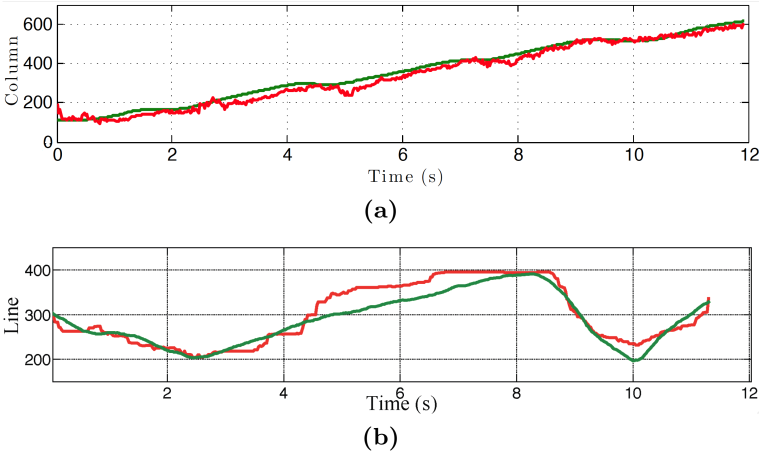
As demonstrated in § 3.4, the DRR is a reliable cue for distance estimation *********and the DRR-SCM cue, proposed by** [Hioka et al., 2010, 2011] **and presented in § 3.3.3 for a binaural application provides reliable distance information*********. Thus, for distance estimation, this technique is employed in a combination with a NN pattern recognition that estimates the distance. *********Hioka's method is applied to provide DRRs on multiple frequency bands********* for each time frame. Then, for each frame, a code vector is composed by a concatenation of the obtained frequency-dependent DRRs, this code vector enters the NN and the distance is output, similarly to the azimuth estimation case. But in this case, and since all the inputs composing the input vector are of the same physical nature, the NN is regularly connected. It has one hidden layer and it is trained with the back-propagation algorithm. The number of DRR frequency bands and thus the dimension of the input vector will be investigated in § 4.1.4.

4.1.3 Reverberations effects dimming

Reverberations highly affect the performances of acoustic cues-based azimuth estimation approaches. The classical approaches that are used to reduce their effects rely on selecting only the frames that are well located and energetic enough that they can be considered as dominated by signal onsets [Heckmann et al., 2006]. This implies neglecting the following or preceding frames that can be considered as damaged and unusable. Similarly, other approaches try to measure the "coherence" between left and right signals: frames that are highly coherent are only supposed to be useful [Liu and Wang, 2010], [Faller and Merimaa, 2004]. These methods, while keeping good features, cause data losses. It is of high interest to use the largest possible amount of the available data, even though a tradeoff between present information reliability and quantity is to be made. In this work, a different operation is applied, it reduces the cue fluctuations induced by the echoes and reverberations without reducing the number of the computed codevectors. Thus, smoothed and less fluctuating cues with less chances of taking values that correspond to positions different from the actual position are provided to the NN. This approach is applied, to the detriment of assuming that the sound source is not moving. Smoothing the distance cues over time can also improve distance estimation performances and is thus applied. This operation is applied as follows: for each time frame, a new feature vector is a weighted sum of the codevectors belonging to a certain surrounding of it. The biggest weight is associated to the current vector and the weights decrease linearly as the corresponding vectors get further: for the vector of the j^{th} frame, a new azimuth cues vector SV_j^{az} is obtained as

$$SV_j^{az} = \frac{1}{(n_v+1)^2} \sum_{l=j-n_v}^{l=j+n_v} (- \mid l-j \mid + n_v + 1) V_l^{az}. \tag{4.1}$$

where V_l^{az} is the l^{th} vector, and n_v is the number of vectors taken before and after the vector j to compute SV_j^{az}. The same applies for distance cues vectors. Note that the computation of the smoothed vector j requires to know the values of a number of vectors with indexes $l > j$, reaching up to the vector $j + n_v$. This imposes a latency of n_v frames on the system operation, which, if not tolerable, can be encountered by taking only frames of index $l \leq j$ in



(a)

(b)

Figure 4.20: Experiments: estimation results, predicted dimensions as a function of time. (a) columns, (b) lines.

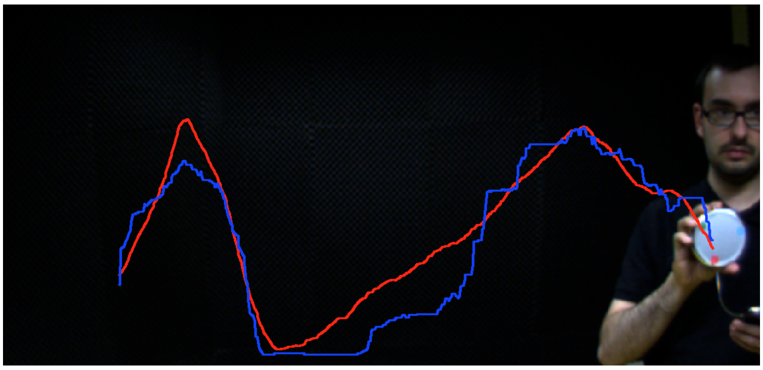


Figure 4.21: Experiments: a truncated view of an image taken by the camera. It shows a predicted trajectory made by the sound source (blue) and the corresponding real trajectory (red).

the interaural cues provided very satisfying results for column estimation, output energies from a set of cochlear filters allowed to efficiently determine the line coordinates. Tests in more complex situations including reverberations and noises are to be made in order to further adapt and evaluate the approach. But the obtained results are encouraging to further proceed in the multimodal information treatment and to model new approaches of audio-visual processing.

4.4 Conclusions and prospects

This chapter first presented an azimuth and distance estimation system. ******* **The used cues were judged to provide the best position discrimination among multiple previously presented techniques** *******. Azimuth estimation relied on interaural cues computed on multiple frequency bands, and the extracted distance cues are also frequency-dependent direct-to-reverberant energy ratios. Neural networks exploited the extracted cues for both tasks. Simulated and real recorded data have been established for evaluating the approach under the constraints of sound reflections and changing receiver positions. First, a careful study was performed in order to set the values of important parameters of the system, then evaluations were made. Results showed very good localization capabilities with a sensitivity to certain factors, notably the environment's acoustic conditions. Later, an alternative visio-

Chapter 5

Conclusion

This manuscript presented works conducted within a PhD thesis enrolled in the domain of artificial audition in the humanoid robotic context. The focus in this thesis was mainly put on the task of sound source localization, brought to the fore by multiple motivations, including its effect on speaker recognition that was also treated. All the conducted works were made in the binaural context, i.e. using signals received in the ears of the robot human-like head. The object was to be located in this biomimetic binaural context, and to obtain the best possible performances at the same time. Thus, there hasn't been a precise reproduction of the processing taking place in the human auditory system, but implementation of computational models of certain parts of this system that were proved to be well used and beneficial. Moreover, the audition performed here can be described as "passive" or "not active", not being explicitly linked to changes occurring on the robot behavior.

In their time sequencing, works started with speaker recognition. The system concatenated MFCC codevectors obtained using the signals of the left and right ears and fed them to a GMM set that modeled the speakers identities. Relying on two signals, with differences related to the speaker's position, this system of binaural speaker recognition was proved to be robust to noises, but sensitive to the speaker position. It provided better performances than a monaural approach implementing the same computational steps. A proposed solution to this sensitivity was to include data extracted from multiple speaker positions in the system training dataset. After speaker recognition, sound source localization has been addressed. A further motivation for this was the fact that the position-related information present in the binaural signals can affect the performances of binaural systems performing other tasks, like the presented speaker recognition system. An observation of the binaural localization state-of-the-art shows that azimuth estimation was mainly addressed, and that although the same cues are extracted for this task, the previously used methods of extracting them present important differences. A first contribution was to implement all these cues on the same database in a statistical quality evaluation object where the cue quality is reflected by its ability to separate different source positions. The conclusions of this study were used in the design of an azimuth and distance estimation system using respectively *********the best found state-of-the-art frequency-dependent interaural cues and DRR computation techniques*********. For both tasks, the cues were exploited by NNs to obtain the azimuth and the distance. The system's robustness to reverberations in multiple aspects was addressed. Azimuth estimation and distance estimation, with the latter relying on a straightforward estimation of the present