TOWARDS A SYSTEMATIC STUDY OF BINAURAL CUES

Karim Youssef, Sylvain Argentieri, and Jean-Luc Zarader

Abstract— **Sound source localization is a need for robotic systems interacting with acoustically-active environments. In this domain, numerous binaural localization studies have been conducted within the last few decades. This paper provides an overview of a number of binaural localization cue extraction techniques. These are carefully addressed and applied on a simulated binaural database. Cues are evaluated in azimuth estimation and their discriminatory effectiveness is studied as a function of the reverberation time with statistical data analysis techniques. Results show that big differences exist between the discriminatory abilities of multiple types of cue extraction methods. Thus a careful cue selection must be performed before establishing a sound localization system.**

Keywords — Robot audition, binaural cues, sound localization, sound processing.

I. INTRODUCTION

Socially interacting robots are becoming more and more interesting and conceivable as partners in human everyday life. Particularly, these robots require sound processing abilities that allow them to detect and separate sounds, recognize sound contents and importantly, localize them. This brings to the fore the sound source localization ability as being one of the major problems for hearing robots. In this context, the last decades witnessed progresses in localization technologies, from microphone arrays to the relatively new field of binaural hearing.

Binaural audition is an emerging biologically-inspired and low-complexity sound processing domain. Relying on signals captured by two ears of a robot, binaural systems try to imitate the human auditory functions that are still hard to reproduce. Most of these systems extract interaural cues that are mainly Interaural Time Difference (ITD), Interaural Level (or Intensity) Difference (ILD or IID) and Interaural Phase Difference (IPD). These cues are used for direction estimation, and more particularly, in the azimuth dimension [13], [15], [21]. In this field, we have recently proposed in [23] some accurate methodologies for estimating the azimuth and elevation of a sound source based on monaural and binaural cues. Whether a localization system aims at estimating the source azimuth, elevation or distance, the same used auditive cues are computed in a lot of different ways and contexts. Then a question arises: what is the best extraction technique for each cue? In an attempt to provide an answer relying on an analysis of the cues themselves, this paper presents a low-level positional discriminatory statistical analysis of multiple techniques. It first provides an overview of sound source localization studies, human auditory system models and methods used to extract acoustical cues and their parameters. It shows that some substantial differences exist between them and thus presenting and comparing them is important. Later, some of them are implemented and applied on a database simulating a realistic sound emission-reception case where multiple levels of reverberations are present. Indeed, realistic environments include the effects of reverberations on the signals, which can not be neglected for artificial auditory systems. Thus this study computes the acoustic cues corresponding to multiple reverberant environments and performs a low-level positional discriminatory ability signal analysis.

The paper is organized as follows: section II provides a review of artificial auditory systems, and azimuth-related cue extraction techniques. Section III presents the dataset established to evaluate these techniques, and the analysis metrics and results. Finally, a conclusion ends the paper.

K. Youssef, S. Argentieri and J.-L. Zarader are with UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005, Paris, France and CNRS, UMR 7222, ISIR, F-75005, Paris, France. E-mail: lastname@isir.upmc.fr

II. LOCALIZATION SYSTEMS: A REVIEW OF AUDITORY SYSTEM MODELS AND CUE EXTRACTION STRATEGIES

Most of the already proposed binaural localization systems in the literature mainly follow the successive steps represented in Figure 1. These auditory system modeling steps are discussed in the first subsection. Next, a binaural cue extraction algorithm must be specified in order to extract some features which will be then used to perform the localization. These algorithms depend on multiple parameters, like time framing, frame durations and overlap, frequency intervals and number of channels. Most of the existing approaches are mainly concerned with the same type of cues, while the ways they are extracted are sometimes very different. This is the reason why a careful review, definitions and comparisons of azimuth specific cues are respectively proposed in the second subsection. This overall review constitutes the first contribution of the paper. Third, a localization method is applied, resulting in source azimuth, elevation and distance estimation, denoted respectively as $\hat{\theta}$, $\hat{\phi}$ and $\hat{r}$. Multiple algorithms for the localization problem exist; one can cite learning approaches like Gaussian models [13] or other approaches using geometrical relations between the positions and the extracted cues [15]. This paper does not discuss this last step, as it only reviews azimuth cue extraction techniques. Finally, a conclusion ends this review section.

A. Auditory System Modeling

Any wavefront reaching the ears is modified successively by the effects of the outer and middle ears, and then by the cochlea inside the inner ear. Biologically-inspired binaural

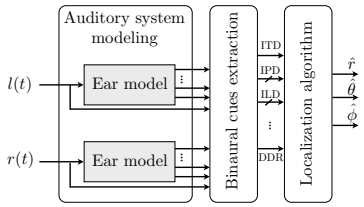

Fig. 1. A typical sound source localization system.

systems are summarized in Figure 2. Multiple models have been already established, with substantial differences existing between them. Generally, the first step consists in reproducing the effect of the outer and middle ears by applying a bandpass-like filter [20]. Then a frequency decomposition is performed, trying to mimic the effect of the basilar membrane inside the cochlea. This frequency analysis can be simulated by applying a gammatone filterbank to the signals [13], [17]. Then, the haircells transduction process is modeled. According to the literature, this last step can be implemented through various approaches, most of them relying on a rectification of the signal followed by its compression. As an example, [5] modeled this overall process as a series of band pass filtering, spectral decomposition, AM-demodulation, A/D conversion and compression. Another frequency decomposition is proposed in [4] where low, intermediate and high frequency domains are defined. Indeed, according to these frequencies, the human auditory system is able to exploit the signal's envelope (higher frequencies) and/or its fine structure (up to 1.5kHz frequencies) [5]. In this field, [8] proposed a system that models the neural transduction as follows: envelope compression (power of 0.23), half-wave rectification, squaring and finally fourth order low-pass filtering with a cutoff frequency of 425Hz.

B. Binaural cues for azimuth Estimation

Once the left and right temporal signals, denoted respectively as $l(t)$ and $r(t)$, are eventually modified by the aforementioned ear model, auditory cues must be extracted. Most of the approaches only focus on azimuth estimation, while dealing with the classical interaural difference cues, namely the ITD, the IPD and the ILD. But while everybody agrees on the information they capture, a lot of very different techniques are used to evaluate their values. These are summarized in the forthcoming subsections. In all the following, all the cues are evaluated on a discrete time-window/frame basis. The two continuous left and right signals are first

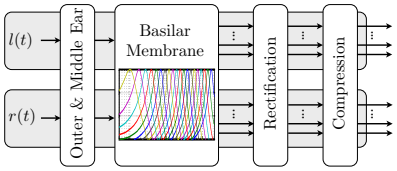

Fig. 2. Ear model: from the raw signals to their multiband frequency representation.

sampled. Their respective discrete values are denoted $l[n]$ and $r[n]$. Then, each signal is decomposed in successive rectangular windows lasting N samples each, thus conducting to N/f_s-length time windows, with f_s the sampling frequency. For convenience, the index of the considered time-window is discarded in all the following notations.

1) Interaural cues extraction from the temporal signals: In this context, ITD and ILD are directly extracted from the two left and right temporal signals for each time frame.

a) ITD: standard cross-correlation (Std-CC): The ITD represents the time required by a wave emitted from a source position to travel from one ear to another. It can be evaluated by using the classical intercorrelation function

$$C_{lr}[m] = \sum_{n=0}^{N-m-1} l[n+m]r[n]. \quad (1)$$

Then, the ITD comes as ITD $= T_s \arg\max_m C_{lr}[m]$, with $T_s = 1/f_s$. This straightforward ITD estimation is very commonly used. Importantly, the ITD then comes as a multiple of the sampling period T_s, thus limiting its resolution. One solution consists then in interpolating the cross-correlation function with polynomial, logarithmic or sinc functions. In this vein, one can cite [12], performing a quadratic interpolation. Also, [8] used a similar cross correlation which is obtained for each sample on a sliding decaying window.

b) ITD: generalized cross-correlation (GCC): Depending on the signal of interest, the standard cross-correlation is known to exhibit not so sharp peaks. A solution consists then in using the generalized cross-correlation, and its well-known PHAT (PHAse Transform) weight function, defined as

$$GC_{lr}[m] = \text{IFFT}\Big(\frac{L[k]R^*[k]}{|L[k]||R[k]|}\Big), \quad (2)$$

where $L[k]$ and $R[k]$ represent respectively the Fourier transforms of $l[n]$ and $r[n]$ obtained through a classical FFT, with $k \in [0, N-1]$ the frequency index[1], and * represents the conjugate operator. For instance, PHAT-GCC is used in [10] to determine the azimuth and discriminate multiple talkers. Note that the aforementioned ITD resolution problems still exist when using GCC, and the same interpolation-based solution can be exploited to improve the estimation resolution.

c) ILD: standard energy ratio (Std-ILD): The ILD, mainly caused by the shadowing effect of the head, represents the intensity difference between the two perceived signals. As such, its definition (in dB) comes as

$$ILD = 20\log_{10}\Big(\frac{\sum_{n=0}^{N-1} l[n]^2}{\sum_{n=0}^{N-1} r[n]^2}\Big). \quad (3)$$

2) Interaural cues extraction from a frequency-dependent analysis: The previous ITD and ILD definitions do not provide any frequency-dependent cues. But ITD and ILD are known to verify the *Duplex Theory* [16], and are therefore respectively dedicated to low and high frequencies. So,

[1]Note that theoretically, $GC_{lr}[m]$ should be computed using a 2N-1 points FFT after zero padding.

a frequency dependent analysis must be performed. But two approaches could be envisioned: on the one hand, a pragmatical engineering-based FFT decomposition can be exploited. But such a Fourier analysis provides at least $N/2$ relevant frequency bins, and thus highly redundant frequency information. So, a mean computation step is often introduced in the literature. On the other hand, human-like frequency decomposition using $K \ll N/2$ gammatone filters can be used (see Figure 2). These two approaches are presented in the following subsections.

a) IPD: spectra angles difference (FFT-IPD): The IPD cue is directly linked to an ITD value related to a specific frequency bin f, with IPD$= 2\pi f$ITD. It can be easily computed from

$$IPD[k] = \arg(L[k]) - \arg(R[k]). \tag{4}$$

As a result, $N/2$ relevant IPD values (for frequencies ranging up to about $f_s/2$ Hz) are extracted from the two signals, thus resulting in a very high-dimensional cue. As a solution, mean computations can be introduced. For example, the IPD can be computed on each frequency bin according to Equation (4), and then averaged over K frequency intervals. This approach will be referred to as FFT-IPD-MEAN1 in the following. Another approach could consist in defining the IPD as the phase difference over the spectra means also computed on K frequency intervals (FFT-IPD-MEAN2 method). Such IPD computations were applied in [15] by addressing the phase unwrapping problem. The same approaches are used in [3], [19], [14], based on 16ms, 32ms and 64ms frame-lengths respectively. Interestingly, [19] proposed to work on frequencies ranging up to 8kHz, while [14] used 43 rectangular channels spanning from 73Hz to 7.5kHz.

b) ILD: spectra magnitude ratios (FFT-ILD): Following the same line as IPD extraction, ILD can be directly defined as

$$ILD[k] = 20\log_{10}\frac{|L[k]|}{|R[k]|}. \tag{5}$$

The same remarks concerning mean computations still apply, thus defining the mean ILD over frequency bands (ILD-FFT-MEAN1), or the ILD computed on the basis of spectra means (ILD-FFT-MEAN2). These ILD definitions are exploited for instance in [15], [19], [14]. A small variation is proposed in [9], where a normalization of the intensity difference with respect to the total intensity in both channels is introduced.

c) ITD: gammatone filters (GAMMA-ITD): Another approach to frequency analysis consists in mimicking the frequency decomposition inside the cochlea (see §II-A). This is mainly performed through K gammatone filters whose center frequencies $f_c[k]$, $k \in [1, K]$, and bandwidths are related to the ERB scale. As a result, K temporal signals –respectively referred to as $l^{(k)}[n]$ and $r^{(k)}[n]$– are available on the left and on the right channels. As a result, cross-correlation operations between these signals can be performed so as to estimate the ITD as a function of the frequency, according to

$$C_{lr}^{(k)}[m] = \sum_{n=0}^{N-m-1} l^{(k)}[n+m]r^{(k)}[n], \tag{6}$$

where $C_{lr}^{(k)}[m]$ represents the inter-correlation computed with the two left and right signals originating from both k^{th} Gammatone filters. Then, in the same vein as §II-B.1.a, the ITD comes as $ITD^{(k)} = T_s \arg\max_m C_{lr}^{(k)}[m]$. This approach has been used in [21] and [22], by adding a normalization of the cross-correlation by the left and right energies product square root. In these works, the time frames were lasting 20ms with 50% overlap, and the filterbank had $K = 128$ filters with center frequencies ranging from 50Hz to 8kHz. Identically, [13] proposed an exponential interpolation allowing the improvement of the ITD resolution. The same frame duration is used here, but the considered frequency decomposition is performed with $K = 32$ filters whose center frequencies spread from 80Hz to 5kHz.

d) ILD: gammatone filters (GAMMA-ILD): Following the same line, ILD (in dB) can also be computed for each gammatone filter thanks to

$$ILD^{(k)} = 20\log_{10}\Big(\frac{\sum_{n=0}^{N-1} l^{(k)}[n]^2}{\sum_{n=0}^{N-1} r^{(k)}[n]^2}\Big). \tag{7}$$

This definition is exploited in [13], [21], [22] in order to obtain ILD values as a function of the frequency.

C. Conclusion

We have proposed in the previous subsection a careful review of auditory cue extraction techniques. From this state-of-the art, a question arises: *how and on what basis should the auditory cue extraction method be chosen?* In other terms, since very different algorithms to obtain the same binaural cue exist, which one is the most appropriate to a specific application? A first natural answer is to choose the one offering the best performances for the aimed task. Then, a natural evaluation metric of auditory cues could be defined, like the localization precision. This is of course a highly relevant metric, but the results are also highly dependent on the used algorithms (models, classifier type, etc.) On the opposite, proposing a kind of low-level, signal-based metric could also give more insight in the appropriateness of a specific cue regarding more general frameworks in robot audition. To our knowledge, such a study has not so far been proposed and will be investigated in the following section. Importantly, this paper mainly focuses on the effects of the reverberations on binaural cues.

III. A SYSTEMATIC STUDY OF BINAURAL CUES

This section aims at defining a signal-based metric for the evaluation of the various auditory cues defined in §II. Importantly, this study must be performed in realistic environments that robotic platforms have to face. This is made possible thanks to the simulation of reverberant environments trough a dedicated MATLAB toolbox. This software and its exploitation will be described in the first subsection. Then, auditory cues will be evaluated in terms of their ability to effectively discriminate multiple sound source positions, on a low/signal-related level. The data analysis technique used in this paper to perform such a study will be explained in the second subsection. Finally, analysis results –regarding azimuth cues– are provided in a third subsection.

A. Generation of the signals database

The forthcoming analysis results are obtained after an offline analysis of the binaural cues thanks to the use of Roomsim [6], a software dedicated to the simulation of the acoustics of a simple shoebox room. Using this MATLAB toolbox, a database simulating multiple acoustic conditions and source-receiver relative positions has been established. Roomsim relies on the images method [2] to generate Binaural Room Impulse Responses (BRIRs), on the basis of anechoic Head Related Impulse Responses (HRIRs) provided by the CIPIC database [1]. In all the following, a $L \times l \times h = 5 \times 4 \times 2.75$m room, with acoustic plaster walls, wooden floor and concrete paint roof is used. Humidity has been set to 50%, and temperature to 20°C. The effects of air absorption and distance attenuation are also taken into account. This configuration gives a reverberation time $RT_{60} = 0.1983$s@1kHz. The walls absorption coefficients were then scaled in order to obtain other datasets with a RT_{60} of 0.45s and 0.7s at 1kHz. The simulated head has been located at the position $(L, l, h) = (2, 2, 1.5)$m, while the source has been placed in multiple positions relatively to the receiver. Azimuth angles always vary between -45° and 45° with a 5° step (thus producing 19 different azimuth values). Distances vary between 1m and 2.8m with a 0.45m step (so that 5 different distances are considered). And for the present study, the source elevation is set to $0°$.

B. Theoretical definition of a signal-based metric

As already mentioned in §II-C, the localization cues effectiveness will not be evaluated in terms of localization errors, since these errors are highly dependent on the used localization/classification techniques. Instead, the proposed analysis is made on the different cues definitions themselves, i.e. their position-dependent dispersions and thus in discriminative abilities. For that purpose, we postulate the use of the Wilks' Lambda metric together with the Linear Discriminant Analysis (LDA) approach, which are both depicted in the following.

1) Theoretical foundation: The dispersions of the set of M (possibly multi-dimensional) features, corresponding to M time frames, can be evaluated through the following successive computations [7], [11]. The dataset is first split to L groups, with m_l the number of features belonging to the l^{th} group, $l \in [1, L]$.

a) Intragroup dispersion (or Within-group dispersion): the intragroup dispersion of the l^{th} group is described by its covariance matrix W_l. The overall intragroup dispersion matrix W for all the data is then defined as

$$W = \frac{1}{M} \sum_{l=1}^{L} m_l W_l.$$

b) Intergroup dispersion (or Between-groups dispersion): the dispersion between different groups is reflected by the intergroup dispersion matrix B defined by

$$B = \frac{1}{M} \sum_{l=1}^{L} m_l (\mu_l - \mu)^T (\mu_l - \mu),$$

where μ_l is the center of the l^{th} group, and $\mu = \frac{1}{L} \sum_{l=1}^{L} \mu_l$.

c) Total dispersion: the total dispersion of the dataset is finally obtained by the total covariance matrix T [11]: $T = B + W$.

2) Wilks' Lambda: Wilks' lambda is a statistical tool that can be used to measure group centers separation. In our case, the Wilks' Lambda, denoted $\wedge$, will be used to estimate the discriminatory ability of a set of auditory cues that can be separated into multiple positional groups. This measurement is defined as being the ratio between the intragroup dispersion and the total dispersion of all the data [7], [18], i.e.

$$\wedge = \frac{\det(W)}{\det(T)}. \tag{8}$$

The smaller the Lambda is, the more discriminant the cue.

3) Linear Discriminant Analysis: LDA aims at describing data that can be separated into multiple groups with discriminant uncorrelated variables. It consists on projecting the data on the basis described by the eigenvectors related to the higher eigenvalues of the matrix $T^{-1}B$ [11]. And a new low-dimensional space which minimizes the intragroup dispersion while maximizing the intergroup dispersion is obtained. Using LDA, a basic classifier can be formed. Data are decomposed into "training" and "testing" data, where training data are used to compute the eigenvectors on which the overall data projection is performed. Only the first two eigenvectors are selected as they capture most of the data variance in this case, and 2D projection is therefore performed. Testing data are then projected on the same 2D space and their minimal euclidean distances to each of the training groups centers specify their group belongings. This gives then a recognition rates performance measure.

C. Cues analysis

We have now recalled all the theoretical background needed to perform the analysis of the auditory cues. As mentioned before, the reverberations effects are carefully addressed and the presented studies provide measures as a function of the reverberation time RT_{60}.

In all the following, data corresponding to the same azimuth angle are taken as belonging to the same azimuth group. So, 19 groups or classes are defined. The auditory cues are all computed in the same conditions, with speech signals lasting approximately 5s and windowed into 23.2ms ($N = 1024$ points) frames, with $f_s = 44.1$kHz. A very basic energy-based Voice Activity Detector (VAD) is then exploited to remove silence frames.

a) The duplex theory: As a first attempt to evaluate if the $\wedge$ is an efficient tool for auditory cues analysis, we propose here to compare the discriminatory abilities of the GAMMA-ITD (defined in (6)) and GAMMA-ILD (see Equation (7)) approaches as a function of the frequency. So in this study, 30 ITDs and 30 ILDs obtained using signals coming from 30 gammatone filters are compared. The resulting $\wedge$ is shown in Figure 3 as a function of the gammatone center frequency. It can be seen that high $\wedge$ values are reached by the ILD in the low index domain,

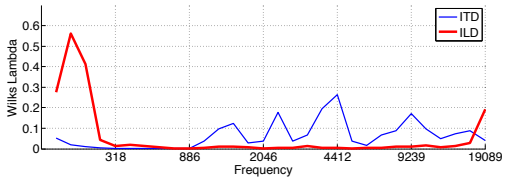

Fig. 3. Wilk's Lambda measures for multiple cochlear filters frequency channels in a 30-filters filterbank.

corresponding to low frequencies. Indeed, ILD values are quite similar for low frequencies since the head effect can be neglected for high wavelengths. Consequently, the ILD is not a discriminative cue for localization in this frequency domain, thus conducting to high Wilks' Lambda values. The same applies to ITD but in the high frequency domain (above about 1.5kHz) because of the phase ambiguity. This effect is known as the duplex theory [16], and is thus "rediscovered" through the proposed approach. This confirms its ability to capture pertinent information regarding cues relevance.

b) FFT-MEAN1 vs. FFT-MEAN2: We have shown in §II-B.2.a and §II-B.2.b that IPD and ILD could be computed with an FFT approach along two strategies. On the one hand, IPD and ILD cues can be computed on each frequency bin according to Equation (4) and (5), and then averaged over K frequency intervals (strategies respectively referred to as IPD-FFT-MEAN1 and ILD-FFT-MEAN1). On the other hand, IPD and ILD can also be defined on the spectra means also computed on K frequency intervals (strategies respectively referred to as IPD-FFT-MEAN2 and ILD-FFT-MEAN2). For this subsection, $K = 30$ adjacent frequency channels are selected between 0Hz and $f_s/2 = 22050$Hz. So for each time frame, 30 ILDs and 30 IPDs are computed. The resulting analysis as a function of reverberation times of both implementations is shown in Figure 4. It can be seen that computing the means of the two cues (MEAN1 approach) is definitely better than computing those of the spectra and then computing the cues (MEAN2 approach). Indeed, the $\wedge$ for this first strategy exhibits lower values, especially for the ILD. The same conclusion is reached regarding the LDA-based recognition rates. So, only the FFT-MEAN1 approach

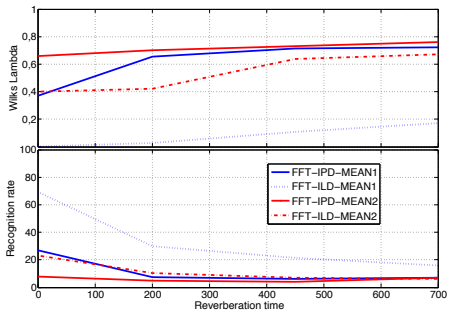

Fig. 4. Wilks' Lambda measures and recognition rates for multiple azimuth FFT-related cues computation techniques as a function of reverberation times.

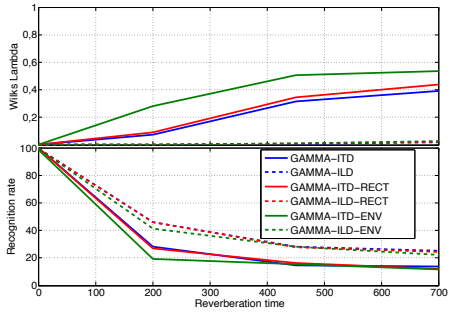

Fig. 5. Wilks' Lambda measures and recognition rates for multiple auditory models-based azimuth cues as a function of reverberation times.

will be considered in the forthcoming comparisons.

c) Auditory models comparison: We have also shown in §II-A that multiple hair cells transduction models exist. 3 of them are compared in this subsection, with the same gammatone filterbank made of $K = 30$ filters covering frequencies of up to 22050Hz. The first approach computes ITD/ILD directly on the original left and right signals (GAMMA-ITD and GAMMA-ILD strategies, see §II-B.2.c and §II-B.2.d). The second approach consists in adding a halfwave rectification combined with a square-root compression step to the previous one (GAMMA-ITD-RECT and GAMMA-ILD-RECT strategies). Third, cues are computed using Bernstein's model [4] (see §II-A): it consists in an envelope compression, half-wave rectification, squaring and finally fourth order low-pass filtering with a cutoff frequency of 425Hz (GAMMA-ITD-ENV and GAMMA-ILD-ENV strategies).

Results of the $\wedge$ and classification rates as a function of the reverberation time are exhibited in Figure 5. It can be seen that the first strategy (i.e. no hair cell model) is the most discriminant in terms of azimuth estimation, while Bernstein's model surprisingly appears to be the least discriminant one. But one has to keep in mind that this last model is assumed to capture what is really happening at the inner hair cells level, while it seems not to be the ideal candidate for an artificial sound source localization system. The human capabilities, although being fascinating for acoustics-related tasks, still seem to have limitations and appear to not use all the possible information contained in the auditory signals. As a consequence, a binaural system designer might have to choose whether he wants to model what is happening in the human auditory system, or to disregard these steps and get better discriminatory performances.

d) Overall comparison: Having studied the FFT-based cues and some of the possible hair cells models, it is now possible to perform a more general comparison between all the auditory cues definitions introduced in §II-B. Figure 6 exhibits the $\wedge$ values and recognition rates for these multiple coding methods as a function of the reverberation time. First, it can be seen that the monodimensional cues, i.e. ITDs and ILDs computed on the two original signals without

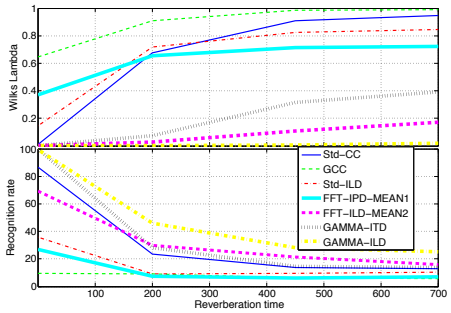

Fig. 6. Wilk's Lambda measures and recognition rates for multiple azimuth-related cues computation techniques as a function of reverberation time

any frequency analysis steps, are the least discriminant, especially in the presence of reverberations. This is definitely not surprising, since considering directly the raw signals does not allow to benefit from the frequency spreading of the reverberation effects. So it appears that considering frequency dependent cues is essential when working on sound localization. The second interesting result is related to the two possible frequency analysis approaches, i.e. FFT vs. gammatone filterbank. Figure 6 shows that ILD/IPD/ITD computed with gammatone filterbanks have the smallest $\wedge$ values and the highest recognition rates with increasing reverberation times. Since gammatone filters frequency intervals are based on the ERB scale, while FFT-based cues are computed over equal adjacent frequency bands, the energy distribution along frequencies highly differs between the two strategies. Noticeably, gammatone filters bandwidth is larger in higher frequencies. But it is known that the reverberation energies are smaller for this same frequency domain, thanks to the classical absorption frequency patterns of the materials classically used in buildings. This might explain why the gammatone filterbank provides the best frequency analysis in terms of separability of the auditory cues.

IV. Conclusion

Multiple sound source localization acoustical cues computation techniques have been reviewed and compared in this paper. Such a study is needed as most localization systems rely on these cues to provide estimations of source positions. These techniques are applied so as to compare them in terms of positions discrimination powers when placed in exactly the same conditions. In this paper, the focus has been put on statistical data analysis as a function of reverberation times. Other influencing parameters and the various elevation and distance cues proposed in the literature will also be heavily studied. Ideally, this work aims at providing a good insight in dynamical cues selection methods, which could provide a meaningful solution to the robust robotic auditory systems problem when operating in the real world.

Acknowledgment

This work was conducted within the French/Japan BINAAHR (BINaural Active Audition for Humanoid Robots) project under Contract n°ANR-09-BLAN-0370-02 funded by the French National Research Agency.

References

[1] V. Algazi, R. Duda, R. Morrisson, and D. Thompson. The cipic hrtf database. *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to audio and Acoustics*, pages pp. 99–102, 2001.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustic Society of America*, 65(4), 1979.

[3] E. Berglund and J. Sitte. Sound source localization through active audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

[4] L. R. Bernstein and C. Trahiotis. The normalized correlation: Accounting for binaural detection across center frequency. *Journal of the Acoustic Society of America*, 100(6), december 1996.

[5] J. Blauert and J. Braash. Binaural signal processing. *IEEE International Conference on Digital Signal Processing*, July 2011.

[6] D. R. Campbell, K. Palomäki, and G. Brown. A matlab simulation of "shoebox" room acoustics for use in research and teaching. *Computer Information Systems*, 9(3), 2005.

[7] A. El Ouardighi, A. El Akadi, and A. Aboutajdine. Feature selection on supervised classification using wilk's lambda statistic. *International Symposium on Computational Intelligence and Intelligent Informatics*, 2007.

[8] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustic Society of America*, 116(5), November 2004.

[9] H. Finger, S.-C. Ruvolo, Paul aznd Liu, and J. R. Movellan. Approaches and databases for online calibration of binaural sound localization for robotic heads. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[10] H.-D. Kim, K. Komatani, T. Ogata, and H. G. Okuno. Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2008.

[11] L. Lebart, M. Piron, and A. Morineau. Statistique exploratoire multidimensionnelle, visualisation et inférence en fouille de données. 2008.

[12] R. Liu and Y. Wang. Azimuthal source localization using interaural cpherence in a robotic dog: Modeling and application. *Robotica, Cambridge University Press*, 28:1013–1020, 2010.

[13] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech and Language Processing*, 19(1), 2011.

[14] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustic Society of America*, 119(1), 2006.

[15] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ILD and ITD. *IEEE Transactions on Audio, Speech and Language Processing*, 18(1), 2010.

[16] L. Rayleigh. On our perception of sound direction. *Philosophical magazine*, 13(74):214–232, 1907.

[17] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, and B. Schölling. Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2006.

[18] G. Saporta. Probabilités, analyse des données et statistique. 1990.

[19] R. J. Weiss, M. I. Mandel, and P. Ellis, Daniel. Combining localization cues and source model constraints for binaural source separation. *Speech Communication*, 53, 2011.

[20] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man and Cybernetics*, 36(5), October 2006.

[21] J. Woodruff and D. Wang. Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7), 2010.

[22] J. Woodruff and D. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[23] K. Youssef, S. Argentieri, and J.-L. Zarader. A binaural sound source localization method using auditive cues and vision. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2012.