2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Using Binaural and Spectral Cues for Azimuth and Elevation Localization

Tobias Rodemann, Gökhan Ince, Frank Joublin, and Christian Goerick

***Abstract*— It is a common assumption that with just two microphones only the azimuth angle of a sound source can be estimated and that a third, orthogonal microphone (or set of microphones) is necessary to estimate the elevation of the source. Recently, using specially designed ears and analyzing spectral cues several researchers managed to estimate sound source elevation with a binaural system. In this work, we show that with two bionic ears both azimuth and elevation angle can be determined using both binaural (e.g. IID and ITD) and spectral cues. This ability can also be used to disambiguate signals coming from the front or back. We present a detailed analysis of both azimuth and elevation localization performance for binaural and spectral cues in comparison. We demonstrate that with a small extension of a standard binaural system a basic elevation estimation capacity can be gained.**

I. Introduction

Most implementations in robotic sound localization focus on the azimuth angle because of its relevance and the relative ease of estimation. The azimuth angle can be estimated with comparatively little effort using only two microphones. It is often stated that to reliably estimate the source elevation more than two microphones are needed ([1], [2]). Alternatively one can make use of spectral cues ([3], [4]), which are thought to require a special shape of the outer ears (pinnae). In this article we will investigate whether special, simplified pinnae are required to use spectral cues and whether those cues are required to estimate the elevation of a sound source with a binaural system. Binaural cues like the interaural time difference (ITD) and interaural intensity difference (IID) are thought to depend only on the angle along the microphone plane, i.e. the azimuth orientation for most standard robotic systems. This assumption, however, is only valid for idealized heads. Real robots' heads are neither perfect spheres nor are they homogeneously filled. In reality, the asymmetric head shape, the head cover, the internal structure (e.g. cameras or computing hardware), and the directionality of microphones and microphone housing will lead to a small, but measurable elevation sensitivity. In this work we are using our robot head with only two human-inspired bionic ears. We compute binaural and spectral cues and investigate the performance for azimuth and elevation sound localization for sources in the front and the back. We will demonstrate that a binaural system can accurately estimate the 2D position of a sound source.

A. Comparison to related work

A serious problem for spectral-cue-based localization is that the spectral structure of a signal at the ear is the result of a position-dependent modulation on a position-independent source signal. Therefore spectral cues are highly dependent on the source signal [5]. A common solution as used in [6] is to limit source signals to white noise. For our system we want to test whether spectral cues can be used with arbitrary sound signals including speech. A further contribution of our work is the analysis of elevation estimation for sources at lateral positions in contrast to the work of e.g. [7], [8] who focused mainly on the medial plane.

The work of Hörnstein et al. [9] presents an approach similar to ours. They also employ a combination of binaural and spectral cues using human-like pinnae to estimate sound source azimuth and elevation. Their approach differs in that they use binaural cues only for azimuth and spectral cues only for elevation estimation. Also they show no detailed comparison of the two cues. Furthermore, in their article no signals from the back of the head are tested and instead of using a human-like pinnae model they use bended metal stripes similar to those used in [8]. In the latter work the authors developed a set of special pinnae which provided a position specific modulation of the source signal spectrum which can be used for localization. The pinnae provided additional features for the binaural and spectral analysis and localization of the sound signals. Our pinnae contain a similar basic structure (the concha, see Fig. 6), but are more complex in structure.

Binaural cues have already been used to estimate both azimuth and elevation by Martin [6] with binaural cues similar to ours. The author's analysis, however, is based on the HRTFs measured on the KEMAR head, i.e. the effect of measurement noise is not considered and no real sound inputs are used. Furthermore, the author did not employ spectral cues.

II. Binaural and spectral cues

In this section we will shortly outline our sound processing architecture, describing how binaural and spectral cues are computed.

A. Binaural Cues

The binaural cues have already been described in detail in [10], [11]. After acquisition from the binaural microphone system we employ a Gammatone Filterbank (GFB) with Equivalent Rectangular Bandwidth (ERB) [12] to split the signal into separate frequency bands (frequency channels). We then extract the two binaural cues Interaural Intensity Difference (IID) and Interaural Time Difference (ITD). Cues are measured only at signal onsets to reduce the effect of

Honda Research Institute Europe, Carl-Legien Strasse 30, 63073 Offenbach, Germany, Tobias.Rodemann@honda-ri.de

978-1-4244-2058-2/08/$25.00 ©2008 IEEE.

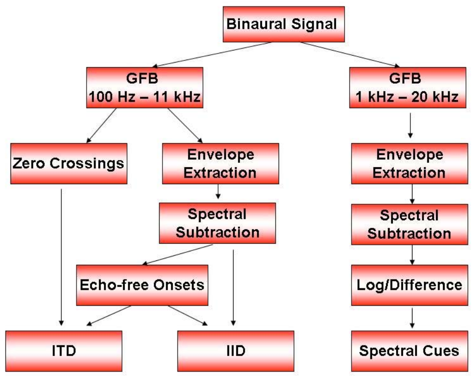

Fig. 1. Comparison of processing chains for the different cues.

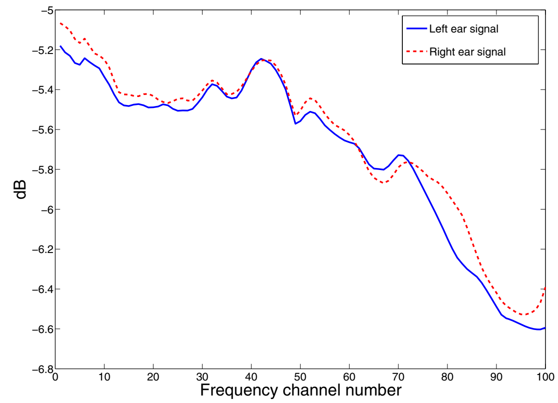

Fig. 2. Spectral vectors of the left and right ear for a source at angle (azimuth = 30°, elevation = 50°). The sound used is the utterance of 'zwei' (German 'two').

echoes [10]. ITD is based on the difference between consecutive zero-crossings of the left and right microphone signal. IID is computed as the difference between the left and right envelope signal divided by the maximum of left and right. In order to reduce stationary noise we employ binaurally synchronized spectral subtraction (see [11]). Binaural cues are extracted for the frequency range between 100 and 11000 Hz in 100 separate frequency channels. See Fig. 1 for a summary of processing chains for the different cues. Note the different ranges of frequencies for binaural and spectral cues.

B. Spectral Cues

Spectral cues are well known from biology. Humans and animals are capable of using the spectral content of incoming sound signals to estimate their position. This ability is based on the direction-dependent frequency filtering effect of the body - especially the pinnae. The result is a selective amplification or attenuation of signal energy in certain frequency bands depending on the direction of the signal.

In our system, spectral cues are computed using the envelope signal after spectral subtraction (as described in [11]). This signal is low-pass filtered and subsampled to 20 Hz. We do not limit the computation of spectral cues to the onset periods which causes spectral cues to be more affected by echoes than binaural cues. Note, however, that echo strength normally decreases with increasing frequency (walls and furniture absorb more high- than low-frequency energy content) so that high-frequency spectral cues should be little distorted by echoes. The spectral cue vector $\vec{s}(k)$ is the energy over all frequency channels (100 channels from 100 Hz to 20000 Hz).

In order to reduce the dependency of spectral cues on the source signal characteristics we compute the difference of the spectral signals in the left and right ear[1] as done also by [9]. Note that our ears are slightly asymmetric between the left and right side (see Fig. 6). A similar approach has been presented in [1], where the same sound was analyzed from two slightly different elevation angles. See Fig. 2 for a visualization of the left and right spectral signal for a short speech phrase presented at 30° azimuth and 50° elevation.

Our experiments yielded that calculating the difference of the signals in logarithmic scale makes the cues more robust. The above-mentioned steps can be given with the following equation:

$$\vec{S}(k) = \log_{10}(\vec{s}_r(k)) - \log_{10}(\vec{s}_l(k)) \tag{1}$$

where $s_{l,r}$ and S are the noise-free envelope signals of left (l) and right (r) ear, and the resulting spectral cue vector, respectively, k denotes the time index. Example spectral vectors for different angles and different sound files are depicted in Fig. 3, Fig. 4 and Fig. 5.

[1] Strictly speaking our spectral cues are binaural cues. To keep the notation from the literature we decided to use the term binaural for the standard, single frequency channel difference cues like IID and ITD and denote cues that operate over the full frequency spectrum as spectral cues.

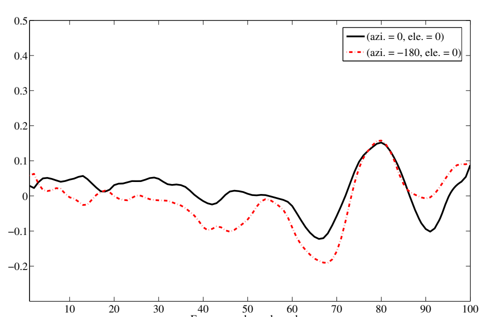

Fig. 3. Comparison of spectral difference vectors for sources in the front (black, solid line) and the back (red, dashed line). The difference vectors are the mean over 29 calibration sounds. Around frequency channel 80 we observe a stable inversion of the relation between the difference vectors.

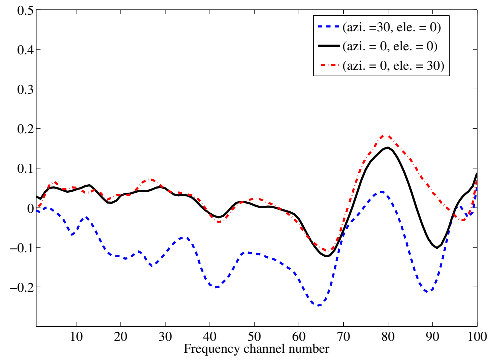

Fig. 4. The spectral difference vectors (mean over 29 different sounds from the calibration procedure) for three different angles. For changes in elevation (compare black and red line) the difference vectors vary little, while for changes in azimuth (black vs. blue line) a more substantial difference is observable, basically due to the IID effect.

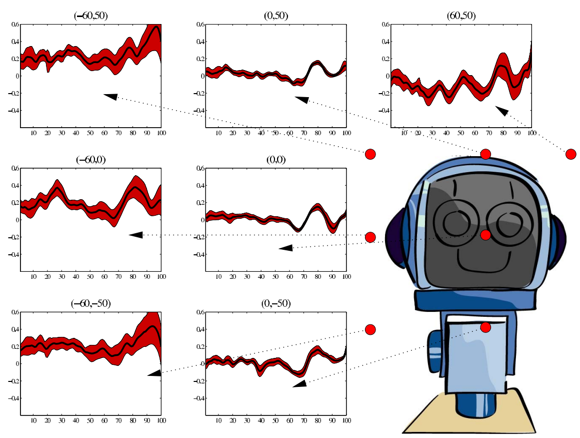

Fig. 5. Spectral vectors for sources from 7 different angles relative to the head. The panels plot mean difference vectors plus-minus one standard deviation over frequency channels. Azimuth angles are -60, 0, +60, elevation angles -50, 0, 50. It is obvious that off-center positions show a higher variation in difference vectors.

III. System

To be able to localize sound sources we first need to calibrate our system by measuring binaural and spectral cues for a large number of positions and different sounds (see below) and using those measurements to calibrate audio-motor maps [13].

A. Cue to Position Mapping

For the localization we generate position estimations for all calibrated positions by using the audio-motor maps to transform cue measurements (i.e. IID, ITD, and Spectral Cues) into position evidence values. This process computes the similarity between reference cue values stored for all positions with the currently measured cues. The process is described in [11] for the binaural cues. Spectral cue evidences are computed by using the *normalized scalar product*. Eq. 2 gives the similarity of the current spectral vector $\vec{S}(k)$ with the reference vector $\vec{S}_{cal}(i,j)$ at position index (i,j). The

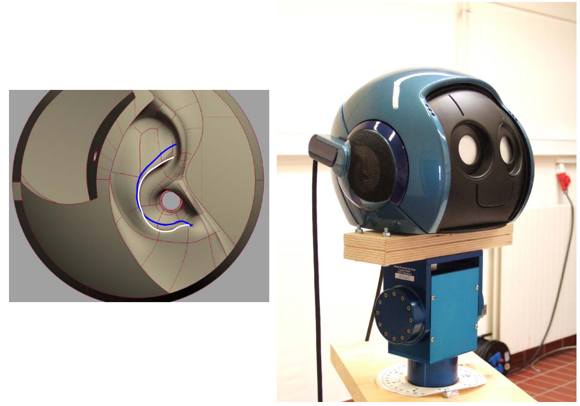

Fig. 6. Left: Bionic ears used in our system. The ears are modeled after the human outer ears. The blue line indicates the concha of the left ear, while the white one corresponds to the right ear's concha outline. Right: Our robot head with two human-like ears on both sides mounted on a pan-tilt element.

indices i,j correspond to the different azimuth and elevation angles in the calibration database.

$$N_{Spectral}(i,j,k) = \frac{\vec{S}_{cal}(i,j) \cdot \vec{S}(k)}{|\vec{S}_{cal}(i,j)| \cdot |\vec{S}(k)|} \tag{2}$$

Position estimations are then integrated over time using the SNR computed by the spectral subtraction as a weighting factor for the current sample. Afterwards binaural and spectral position evidence values are combined. For the combination we are currently using a simple addition with a scaling factor C_S for the spectral cues:

$$N_{Combined}(i,j,k) = N_{Binaural}(i,j,k) + C_S * N_{Spectral}(i,j,k) \tag{3}$$

In our system the scaling factor C_S was set to 100.

B. Hardware and Ear Design

We employ a set of two human-like outer ears custom-made from silicon in a plastic shielding. The microphones are standard Sennheiser DPA 4060-BM omnidirectional microphones. Fig. 6 (left side) visualizes the shape of our ears. The blue and white lines indicate the outline of the concha structure for the left and right pinna, respectively. In addition to providing spectral cues the ears are designed to attenuate signals from the back of the head (specifically fan noise of Asimo) relative to signals from the front. Our tests showed an amplification of signals from the front relative to those from the back of up to 4.5 dB.

The head (neck) motor element is an Amtec Robotics PowerCube070. Due to its ego-motion noise we interrupt sound localization when the speed exceeds a certain value. The combined system is shown in Fig. 6 (right side).

C. Online System and Position Prior

We have also implemented an online version of the system. Integrated position evidence is transformed by a Kalman-filter-based tracker into head steering commands. The online system runs on two standard (multi-core) computers (plus separate systems for sound acquisition and head motor control). The system is developed and runs within the ToolBOS framework described in [14].

As will be discussed later, the performance for elevation estimation depends on the azimuth angle and for lateral positions the precision is not very high. In these situations, the elevation localization shows large errors which often lead the system to falsely look at positions where sources normally do not appear (e.g. below or above the robot). To avoid this type of behavior we introduced a prior on the elevation angle. Effectively, we multiply localization evidence with a position dependent prior information on the likelihood of different elevation values ϕ. This prior is modeled as follows:

$$P(\phi) = \exp\left(-\frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}\right) \quad , \tag{4}$$

with ϕ_0 as the most likely elevation (0° in our case) and $\sigma_\phi = 95°$ as the elevation scaling factor. Note that the prior is applied in an absolute (torso) coordinate system, not in head-relative coordinates.

As an effect, the localization system is given a tendency towards picking the a-priori most likely angle (ϕ_0), especially for highly ambiguous situations, e.g. if the source is in a lateral position. In situations of higher precision and confidence (e.g. the sound source is in a more frontal position) the prior will only lead to a small shift towards the ϕ_0 orientation. Note that the prior was not used for the data reported on in the next section.

IV. RESULTS

We firstly evaluated the system in an offline scenario. Data was recorded in a noisy, very echoic room ($T_{60} = 1100ms$) and some of the 57 sound files in our database were used for calibration and the remaining for testing. Most sound files are short speech phrases but there are also other types of sound (e.g. music, telephone ringing, white noise). In the first calibration scenario 15 calibration files were used (all spoken by the same speaker). The remaining 42 files were used for testing. In another scenario we used half (29) of the sound files for calibration and half (28) for testing, giving us a far better coverage of typical sounds than in the first case. We wanted to test if the size and variability of the calibration set influences the performance of the localization. In the calibration process the sound files were played from regularly spaced angles (azimuth -180° to +180° in steps of 10°, elevation from -60° to 60° also in steps of 10°, the distance was kept constant at about one meter). The test files were produced from the same set of angles. In total we calibrated the system for 36 different azimuth and 13 different elevation angles by storing reference spectral difference vectors and IID plus ITD values which are computed as the mean of measured cue values for all calibration files.

Results are described in Table I for the performance in azimuth localization, in Table II for elevation. Values are given as mean localization errors (difference between true and estimated angle) in degrees and as the percentage of correct responses (within the 10° recording precision). Values are averages over all test sounds and elevation angles. The first column shows the mean error averaged over all azimuth angles, while the second and third column provide separate values for the central (foveal) region of ±30° in the front and back of the robot (second column) and for the remaining azimuth angles (third column). Table III provides results for the rate of front-back confusions. Note that when front and back have been confused (e.g. a target at 0° is localized at 170°) this is counted as a front-back confusion and an additional azimuth error (of 10° in the example case). In Table IV results for the two different calibration scenarios are compared.

cues	azi. error	fovea[-30 30]	rest
Binaural (29)	3.0°(64%)	1.7°(78%)	3.9°(56%)
Spectral (29)	11.0°(49%)	6.4°(63%)	13.9°(40%)
Combined (29)	2.8°(71%)	1.1°(85%)	3.9°(63%)

TABLE I

AZIMUTH ERROR IN DEGREES AND PERCENTAGE OF CORRECT ESTIMATIONS OF AZIMUTH

cues	ele. error	fovea[-30 30]	rest
Binaural (29)	20.0°(39%)	13.1°(54%)	24.4°(30%)
Spectral (29)	19.6°(42%)	13.5°(58%)	23.6°(31%)
Combined (29)	12.3°(54%)	6.8°(73%)	15.8°(42%)

TABLE II

ELEVATION ERROR IN DEGREES AND PERCENTAGE OF CORRECT ESTIMATIONS OF ELEVATION

Our results can be summarized as follows:

- Using only binaural cues it is possible to accurately estimate the azimuth angles for sources both in the front and the back of the head (3° mean azimuth error). The probability of correct elevation estimation is also surprisingly good at about 40% (consider that the chance rate is less than 8%). Looking at front-back confusion we see that binaural cues on their own are correct in more than 85% of all sound files and angles (see Table III).
- As targeted, spectral cues are capable of providing a good estimation of sound sources' elevation angle (19.6° mean elevation error). In this regard they are slightly better than binaural cues. Spectral cues are even capable of estimating azimuth angles, although weaker than binaural cues (11° localization error compared to 3° for binaural cues). Their ability to solve the front-back-confusion problem is as good as for binaural cues.
- The results for the combined cues show that binaural and spectral cues are using at least partially orthogonal information since in combination the results are substantially better in all aspects. We get a very good result of 2.8° azimuth and 12.3° elevation error and a front-back confusion rate of just 7%.
- The localization error and front-back confusion rates

scenario	confusion rate	fovea[-30 30]	rest
Binaural (29)	13.0%	8.0%	16.2%
Spectral (29)	14.8%	8.4%	19%
Combined (29)	7.2%	3.3%	9.8%

TABLE III

FRONT-BACK CONFUSION RATES

scenario	confusion rate	azi. error	ele. error
15 training sounds	10.4%	4.6°	20.7°
29 training sounds	7.2%	2.8°	12.3°

TABLE IV

COMPARISON OF THE RESULTS FOR DIFFERENT SCENARIOS

vary by up to a factor of 10 for different azimuth angles (see Fig. 7). Lowest error rates can be observed directly in front of the robot (0.3° in azimuth and 3° in elevation), while much higher values result for lateral sources (at $-90°$ the mean localization error is 9.4° for azimuth and 23.3° for elevation). Even within a broader foveal region up to 30° from the medial line in the front and back the localization performance is much better than at more lateral positions (up to 50% reduction in localization errors).

- The localization error for different elevation angles is visualized in Fig. 8. For spectral cues we can observe a better performance for the central elevation angles and higher error values for the extreme values of elevation (a kind of fovea effect). In contrast, the performance of binaural cues varies little with elevation angle. For spectral cues we note that the mean localization error for azimuth and elevation is quite similar (especially in the center), while binaural and the combined cues exhibit a substantially lower error for azimuth than for elevation estimation.
- Comparing the two different calibration scenarios we see that the localization performance profits in all aspects from a larger, more representative calibration set. It might therefore be worth investigating the optimal calibration set in more detail.
- The precision of binaural cues has already been evaluated in [11], where we showed that in the foveal region there is an effective resolution of 1° . In order to study the precision of spectral cues around $\pm$ 9° in azimuth and elevation) we recorded a small set of longer (ca. 15 s long) test files for this region. The recording resolution was set at 3°. We now computed the mean localization error of spectral cues. The correct azimuth position was estimated in 84% of all test sounds, for elevation the value is 65%. The mean localization error is 0.6° in azimuth and 1.4° in elevation. The results demonstrate that a sound source that is active for several seconds can be localized with a high precision even by spectral cues. This is also confirmed by the spectral difference vectors (not shown), which vary consistently even for bearings only 3° apart.

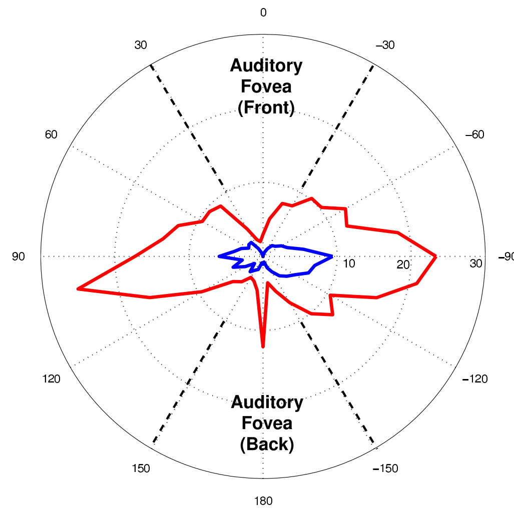

Fig. 7. Localization error as a function of azimuth angle for the combined cues. The (outer) red line indicates the mean elevation error (averaged over all elevation positions and test sounds), the blue (inner) one corresponds to the azimuth error. Within an auditory foveal region between approximately -30 and 30° azimuth the localization error is below 5° for azimuth and less than 10° for elevation.

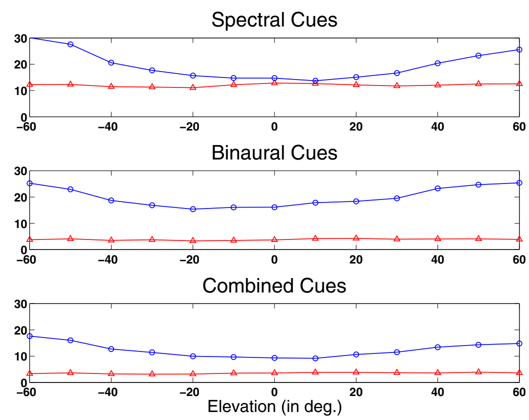

Fig. 8. Localization error (azimuth (red triangle) and elevation (blue circle)) in degrees as function of elevation angle. Values are averages over all azimuth positions and test sounds. The different panels show the performance of spectral cues (upper panel), binaural cues (center panel), and combined cues (lower panel).

We also tested our system in an online scenario. The system was put at various positions in a room different from the one used to record the calibration data. We asked people to interact with the system triggering an orienting motion from all around the robot head. Despite some mislocalization, especially at lateral positions, the system managed to focus on the sound source after a few utterances or other sounds (claps, knocking on table, footsteps).

V. SUMMARY AND OUTLOOK

We have demonstrated in this work that binaural cues can be used to estimate the sound source elevation and to disambiguate front and back. We also showed that spectral cues can be used with human-like bionic ears, even though there is no clear notch- or peak-position to source angle relation as was used in other approaches [8], [9]. Nevertheless using a calibration of the system with a reasonable set of sound data, spectral cues can be used. Our approach shows its full potential in the combination of binaural and spectral cues which, when integrated, show a very good localization performance. Even azimuth localization can profit from the addition of spectral cues.

Although the test database was analyzed offline, the files were recorded in a very noisy and echoic room and we employed a variety of sound signals including speech instead of white noise signals. The described concept was also implemented in a real-time online system using standard computing hardware, where it also exhibited a robust localization behaviour.

In its current form, our system would have difficulties dealing with multiple, simultaneously active sound sources. Especially spectral cues would be affected, since they extend over all frequencies. Therefore, limiting spectral cues to a smaller range of frequency channels might actually improve performance in multi-source scenarios. The general solution, however, would be to separate different sources before localizing them [15].

The advantage of our approach is that with minimal hardware effort (just two microphones on a standard robot head) a sound localization in 2D is possible. All it takes is the extraction of spectral cues and a calibration for different elevation angles. Tests we have conducted with other robot heads (with and without pinnae) indicated that some elevation cues are available even if no special pinnae are used. Furthermore our data suggests that it is possible to get a 2D localization without any spectral cues, purely relying on standard binaural cues.

In an interacting system, e.g. a robot that orients towards sound sources, larger localization errors and especially front-back confusions are perceived very negatively. The high error rates at lateral positions require a careful treatment of raw localization results. In many cases it might be advisable not to orient towards a perceived sound source if the localization data is highly ambiguous. Therefore, we suggest a behavioral strategy for an active robot system that first uses the relatively precise azimuth estimation to bring the sound source into the auditory fovea [16] and then use a second measurement to determine the elevation and the precise azimuth angle. This behavior is induced by the prior mentioned in section III.

Spectral cues could probably be improved in a number of ways, e.g. by optimizing the frequency range over which sounds are compared. The biggest problem in the current approach is that it requires long calibration sessions since in the full 2D space binaural and spectral cues have to be measured. Reducing this calibration effort and making an online adaptation [13] possible are our next targets. We also plan to investigate the potential for distance estimation using the same approach.

VI. Acknowledgments

We would like to thank Antonello Ceravola, Marcus Stein, Benjamin Dittes, and Mark Dunn for providing the necessary hardware and software foundations for our work.

References

[1] D. Gill, L. Troyansky, and I. Nelken, “Auditory localization using direction-dependent spectral information,” *Neurocomputing*, vol. 32–33, pp. 767–773, 2000.

[2] A. Ovcharenko, S. J. Cho, and U.-P. Chong, “Front-back confusion resolution in three-dimensional sound localisation using databases built with a dummy-head,” *J. Acoust. Soc. Am*, vol. 122, no. 1, pp. 489–495, 2007.

[3] M. L. Spezio, C. H. Keller, R. T. Marrocco, and T. T. Takahashi, “Head-related transfer functions of the rhesus monkey,” *Hearing Research*, vol. 144, pp. 73–88, 2000.

[4] M. M. Van Wanrooij and A. J. Van Opstal, “Contribution of head shadow and pinna cues to chronic monaural sound localization,” *Journal of Neuroscience*, vol. 24, no. 17, pp. 4163–4171, 2004.

[5] P. M. Hofman and A. J. V. Opstal, “Spectro-temporal factors in two-dimensional human sound localization,” *J. Acoust. Soc. Am.*, vol. 103, no. 5, pp. 2634–2648, 1998.

[6] K. Martin, “Estimating azimuth and elevation from interaural differences,” in *1995 IEEE Mohonk workshop on Applications of Signal Processing to Acoustics and Audio*, 1995.

[7] M. Kumon, T. Shimoda, R. Kohzawa, I. Mizumoto, and Z. Iwai, “Audio servo for robotic systems with pinnae,” in *Proc. Int. Conf. Intelligent Robots and Systems (IROS) ’05*, Edmonton, Canada, 2005, pp. 885–890.

[8] O. Ichikawa, T. Takiguchi, and M. Nishimura, “Sound source localization using a pinna-based profile fitting method,” in *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, 2003, pp. 263–266.

[9] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, “Sound localization for humanoid robots - building audio-motor maps based on HRTF,” in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*. IEEE, 2006.

[10] M. Heckmann, T. Rodemann, B. Schölling, F. Joublin, and C. Goerick, “Binaural auditory inspired robust sound source localization in echoic and noisy environments,” in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*. IEEE, 2006.

[11] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin, and C. Goerick, “Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping,” in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*. IEEE, 2006.

[12] M. Slaney, “An efficient implementation of the patterson-holdsworth auditory filterbank,,” Apple Computer Co., Technical Report 35, 1993.

[13] T. Rodemann, K. Karova, F. Joublin, and C. Goerick, “Purely auditory online-adaptation of auditory-motor maps,” in *IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2007)*. IEEE, 2007.

[14] A. Ceravola, M. Stein, and C. Goerick, “Researching and developing a real-time infrastructure for intelligent systems - evolution of an integrated approach,” *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 14–28, 2008.

[15] K. Nakadai, S. Yamamoto, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, “A robot referee for rock-paper-scissors sound games,” in *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA), May 19-23, 2008, Pasadena Conference Center, Pasadena, CA, USA*, 2008.

[16] K. Nakadai, H. G. Okuno, and H. Kitano, “Auditory fovea based speech enhancement and its application to human-robot dialog system,” in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, IEEE, Lausanne, Swiss, 2002, pp. pp.1314–1319.