The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

A study on distance estimation in binaural sound localization

Tobias Rodemann

***Abstract*— The position of a sound source is an important information for robotic systems to be extracted from a sound. Of the three spherical coordinates (azimuth, elevation, distance) only the azimuth direction is extracted in most robot audition systems. So far rarely investigated is the issue of estimating the distance between robot and sound source. In this article we describe a study on distance estimation using a binaural robot system in an indoor environment for sounds ranging in distance from 0.5 to 6m. We investigated several proposed audio cues like interaural differences (IID and ITD), sound amplitude, and spectral characteristics. All cues are computed within the framework of audio proto objects. In an extensive experimental setup with more than 10000 sounds we found that both mean signal amplitude and binaural cues can, under certain circumstances, provide a very reliable distance estimation. There was no observable effect of frequency dependent attenuation so that the spectral amplitude cue was only slightly above chance level. We also investigated the loss of precision of azimuth estimation with distance. In contrast to what could be expected, the performance does not severely deteriorate when the system is calibrated for different distances.**

I. Introduction

In robotic sound localization the position of a sound source is often equated with the azimuth angle. This is due to the focus on interaction with humans which tend to be separated by azimuth rather than elevation orientation. Therefore standard sound localization systems are optimized for azimuth estimation via a planar arrangement of microphones. Recently, the estimation of elevation using only two microphones has gained momentum ([1]–[3]). Distance, however, is still largely ignored, although in many robot audition scenarios the behavioural relevance of sound sources decreases with distance. Physics suggests that the measured signal amplitude, which falls off with distance, could be a potential distance cue [4]. Furthermore, the frequency-specific attenuation effect of the air should result in lower amplitudes in the upper frequency range with increasing distance, an effect well-known when hearing distant thunder. We compare the performance of these distance estimation cues with the use of the standard binaural difference cues (interaural intensity difference (IID) and interaural temporal difference (ITD)) extended for distance. We show that a combination of these cues can provide a good distance estimation for a binaural robot.

A. Comparison to related work

In robot audition, distance estimation has so far largely been based on either motion parallax (triangulation) [5], [6] or large scale microphone arrays [7]. Triangulation requires a mobile robot, a continuously active sound source and time to estimate the distance to the source. Larger microphone structures are not mobile and have not been shown to reliably estimate the distance of a sound source that is at a substantial distance from the array's microphones.

Interestingly, animals seem to be able to robustly estimate the distance to a sound source in complex environments using only two ears and without the necessity for ego-motion [8], [9]. It appears that this capacity is based on cues like the ratio of direct-to-reverberant energy or the signal's amplitude.

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

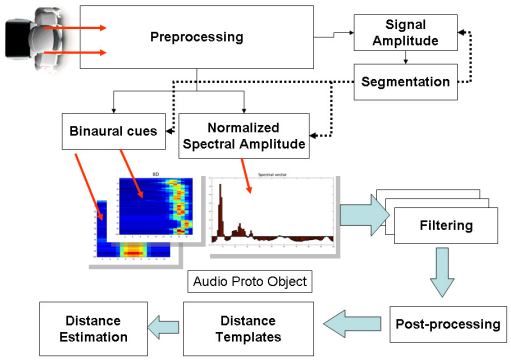

Fig. 1. System architecture (the preprocessing module is described in more detail in Fig. 2).

II. Audio proto objects

In [10] we have introduced the concept of audio proto objects as a mid-level representation of sounds. Basically, an audio proto object is a collection of compressed audio features, that represent the mean characteristics of a sound segment. The system graph for distance estimation in the framework of audio proto objects is shown in Fig. 1. We recorded sounds from two microphones at 48 kHz sampling rate. We employ an auditory preprocessing using a Gammatone Filterbank (GFB) as a model of the cochlea [11]. Based on this signal representation we calculate a number of audio features. The result is a crisp representation of a sound for which we want to estimate the distance. This process has already integrated feature values over all samples of a segment, which substantially reduces the variability of audio features.

A. Audio feature extraction

The feature extraction system is shown in Fig. 2. It is similar to and based on the system explained in [10]. We compute binaural localization cues (ITD and IID), using

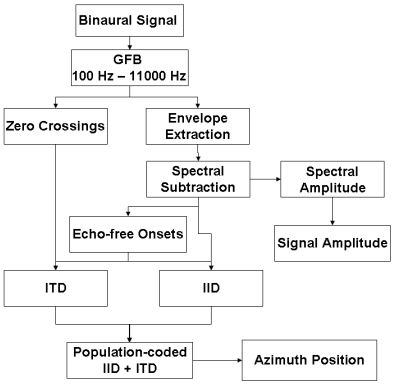

Fig. 2. Sketch of the computation of audio features.

onset-only measurements to reduce the effects of echoes. ITD uses differences in zero-crossings of the cochlea signal, while IID works on the signal envelope after spectral subtraction which removes most of the stationary background noise. The spectral amplitude (the envelope signal for different frequencies) is another candidate cue for distance estimation. We also compute the signal amplitude, which is the sum of envelope values over all frequencies for a single sample.

B. Segmentation process

The segmentation is based on a simple thresholding process working on the signal amplitude. If the signal amplitude rises above a threshold the segment starts, if it falls below the threshold the segment ends. This process is very simple, easy to tune and sufficient in most scenarios. It will obviously fail in the case of several concurrent sound sources. How to deal with this problem (sound source separation) is outside the scope of this paper.

C. Feature compression

The final stage of the proto object generation is compressing the relevant audio features to a representation that is shorter and independent of segment length. We use a simple averaging over time for the signal amplitude and spectral amplitude (separately for all frequency channels). For the binaural localization cues IID and ITD we perform a histogram like averaging process as outlined below. After this stage a set of filter modules removes audio proto objects that are either too short or whose signal amplitude is too weak (this function was not used for the data presented here).

III. Audio cues to distance

A. Signal amplitude

The most obvious cue to distance is the sound amplitude A which falls off with distance d:

$$A = A_0/d \quad , \tag{1}$$

where A_0 is the production amplitude (measured at a reference distance of e.g. $d = 1m$). In our binaural system the amplitude is computed as the sample-wise maximum of left and right amplitude signal. Distance is then computed as:

$$d = A_0/A \quad . \tag{2}$$

There are two problems associated with this approach - this relation is only strictly valid under ideal conditions (as shown in Fig. 5) due to echo effects, and the production amplitude is normally unknown. As a first solution to the latter problem we approximate A_0 by the mean value of all sounds in the training set from a distance of 1 m. The former problem is solved as an extension of the previous idea by measuring average signal amplitude values over all N_C sounds and N_α azimuth angles:

$$A_c(d) = \frac{1}{N_\alpha \cdot N_C} \sum_\alpha \sum_c \quad A^{train}(\alpha, c, d) \quad . \tag{3}$$

$A^{train}(\alpha, c, d)$ denotes a measurement of the mean signal amplitude for a sound c at azimuth angle α and distance d. We note that we generally got slightly better results if the mean was replaced by the median computation. For estimating the distance of a new sound we search for the distance d for which the difference between measured signal amplitude A_m and calibrated (averaged) signal amplitude $A_c(d)$ are minimal:

$$d = \arg\min_d(||A_c(d) - A_m||) \quad . \tag{4}$$

We note that $A_c(d)$ is adapted to the specific environment in the lab. It deviates from the theoretical form only for distances above 3m. We therefore expect only a small improvement from using eqn. 4 rather than eqn. 2. A more substantial benefit would probably result if the range of distances is extended further. In Fig. 5 the theoretical values are compared with measured ones from our training set.

B. Spectral amplitude

We tested the spectral amplitude as a possible cue for distance estimation, although considered to be only effective over larger distances. The spectral amplitude $W(f)$ is the average amplitude in frequency channel f integrated over all samples of the segment. Since it implicitly contains the signal amplitude we have already exploited for distance estimation, we normalize the spectral amplitude $\vec{W}$ to a mean of zero and a norm of one. We then compute a prototypical spectral amplitude vector $\vec{W}_c(d)$ for every distance d as the average spectral vector for this distance over all azimuth angles and all training sounds. The distance estimation for a new sound with a spectral amplitude vector $\vec{W}_m$ is performed as:

$$d = \arg\min_d(\sum_f |(W_c(d, f) - W_m(f))|) \quad . \tag{5}$$

C. Binaural cues

Most sound localization systems are based on measuring ITD and IID to determine the sound position. While some approaches extend to 3D [12], binaural cues are mostly employed for horizontal localization. We could show [3] that due to the interaction with the robot's body binaural cues vary with elevation angle and therefore a combined azimuth

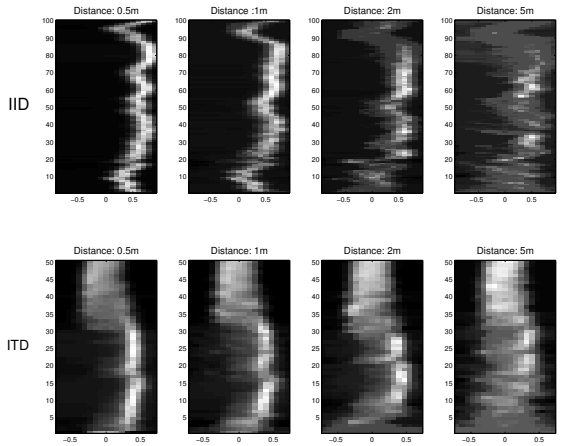

Fig. 3. Example of IID and ITD histograms for a sound at -60 degrees azimuth and different distances. Each graph shows histogram values for different bin centers and frequency channels. For ITD we only show the first 50 frequency channels, since for higher frequencies all ITD values are close to zero.

and elevation estimation is possible using only binaural cues. Our approach for binaural cue-based distance estimation is based on extending the standard horizontal sound localization by calibrating also for different distances. We compare an audio proto object with stored binaural cue histogram templates for all positions (azimuth angle and distance) and take the best matching one (highest scalar product value) as the candidate using both the related azimuth and distance values. In the audio proto object that represents a sound, we collect all measurements of IID and ITD. For every frequency channel f we compute a histogram H_f of binaural cue values (separate for IID and ITD). The histogram bins cover typical values of IID and ITD (in the range $[-0.9, 0.9]$). All histograms are normalized to a mean of zero and a norm of one. In a calibration session we compute these histograms for different positions (in azimuth and distance) as the average values over a number of training sounds. As a result of this procedure we get for every position $p = (d, \alpha)$ the characteristic IID and ITD distributions $H_f^{IID,ITD}(p)$. For an unknown sound, position estimation is performed by comparing cue histograms $H^{IID,ITD}$ for this sound with the histograms of all positions. The similarity S to the template at position p is computed by a scalar product over all frequencies f and histogram bins n:

$$S^{IID,ITD}(p) = \sum_f \sum_n H_{f,n}^{IID,ITD} * H_{f,n}^{IID,ITD}(p) \quad . \qquad (6)$$

The values for IID and ITD are computed separately and then averaged. The similarity value $S(p)$ is directly taken as the evidence for position p. The most likely position of a sound source is the one with the highest similarity value.

IV. Methods

Our recording scenario consists of a robot head mounted on a pan-tilt unit with two ears attached to the sides. The

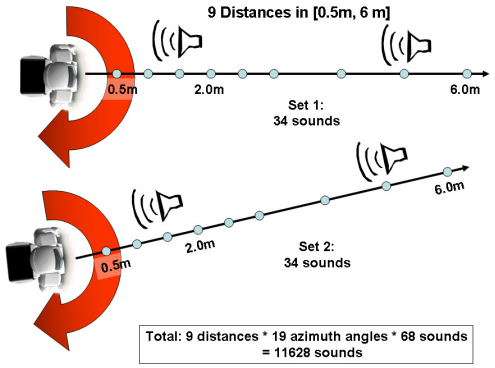

Fig. 4. Outline of the experimental sessions to record training and test sets. Note that the two sets were recorded at slightly different positions of robot and speakers

head is positioned approximately 1 m away from the wall in a typical robot lab environment of dimensions (12 x 11 x 2.8 m) with a substantial amount of echo ($T_{60} = 810ms$). Sounds are generated by a loudspeaker that is positioned in front of the head (azimuth = elevation = 0 degrees). The loudspeaker is put at different distances from the head (from 0.5 to 6 m). Furthermore we horizontally rotate the head between +90 and -90 degrees in order to get a relative change in azimuth position. For each combination of speaker distance and head pan angle we play a number of sound files. These sounds consist of human speech from various speakers, environmental sounds, and music. In total we play 68 different sounds. See Fig. 4 for a sketch of the recording sessions. The sounds were recorded in two separate sessions (Set1 and Set2) with different positions of the speaker in the room and different sound files (34 each). Both training and test sets contain half of the sounds each from Set1 and Set2 and all azimuth angles. The training set was used to compute mean (median) values of audio features (e.g. mean amplitude) for every distance.

A. Correction factors

The measured mean signal amplitudes show a high variability for different sounds that is often larger than the effect of distance. Two factors for this variability we have observed are the source's horizontal position (the measured signal amplitude depends on the relative orientation of the robot's head) and the type of sound. Based on the azimuth localization we can compute a position dependent correction factor for the measured signal amplitude. The azimuth correction factor $C_A(\alpha)$ is inserted into eqn. 4 by replacing:

$$A_m \leftarrow C_A(\alpha) \cdot A_M \quad . \qquad (7)$$

We assume that the azimuth localization system can reliably extract the correct azimuth angle α. Empirical correction factors varied between 0.93 and 1.1. A variation of this approach is to predict the signal's production amplitude. This could be done based for example on the spectral characteristics of the sound. Assuming that such a sound

Cue	mean error	rel. error	correct	conf. near/far
Random	2.5	1.22	11%	9.5%
Amplitude (raw)	1.83	0.55	25.8%	2.7%
*Amplitude (azi)	1.83	0.54	25.9%	2.6%
Amplitude (azi+cat)	0.9	0.17	58.1%	0.02%
*Spectral	2.34	1.0	21%	10.5%
*IID	1.14	0.15	71.3%	1.9%
*ITD	1.61	0.43	44.4%	2.9%
Combined	0.36	0.27	77%	0.12%

TABLE I

DISTANCE ESTIMATION PERFORMANCE . FOR SIGNAL AMPLITUDE WE COMPARE THREE DIFFERENT APPROACHES: USING THE *raw* AMPLITUDE MEASUREMENTS AND EQN. 4, APPLYING AN AZIMUTH CORRECTION (*azi*) AS IN EQN. 7, AND COMBINED AZIMUTH AND CATEGORY CORRECTION AS IN EQN. 8. THE RESULTS FOR A COMBINATION OF ALL CUES (MARKED WITH A STAR '*') ARE SHOWN IN THE LAST ROW.

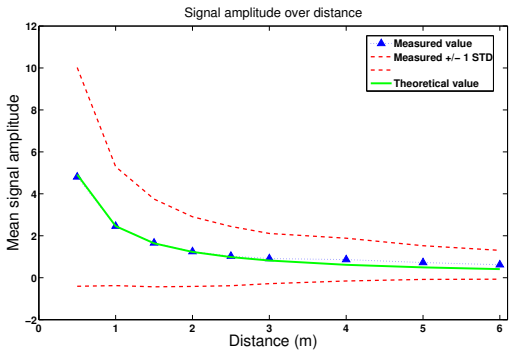

Fig. 5. Mean values (plus/minus 1 standard deviation, computed on the training set) and theoretical values for signal amplitude over distance. Note the small deviation between theory and experiment for distances above 3 m. Values are averages over azimuth angles and sounds.

categorization is possible, the distance estimation can now be based on the concept of 'familiar distance'. For this investigation we tested the performance gain when A_0 is further modified by a sound-type specific correction factor C_C:

$$A_m \leftarrow C_C \cdot C_A(\alpha) \cdot A_m \quad . \tag{8}$$

C_C was derived from the training set by computing the average measured signal amplitude for the specific type of sound relative to the mean over all sounds. Since we don't try to estimate the sound type in this work, the results using this correction factor can be assumed to represent a best case scenario. We determined correction factors in the range of 0.1 to 5.9.

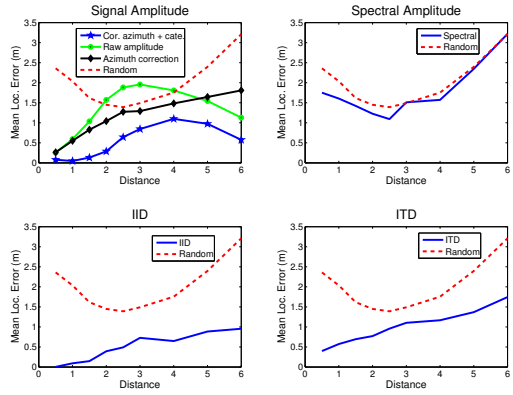

Fig. 6. Mean localization error averaged over all azimuth angles and sounds in the test set for the different localization cues at different distances.

V. RESULTS

We examined the four distance estimation cues (signal amplitude, spectral amplitude, IID, and ITD) with a number of measures: the first one is the mean distance estimation error (averaged over all azimuth angles, distances, and sound files). We also measured the relative error (estimation error divided by true distance). We further computed the percentage of correct estimations (error = 0), for which the baseline is 1/9 (11%). Finally we looked at the percentage of severe mislocalizations (termed near/far confusion) where the localization error exceeds 4 positions (e.g a sound at 0.5 m estimated as 3 m and further or a sound at 6 m localized at 2 m or less). The results are shown in table I. A more detailed comparison for different distances and different cues is shown in Fig. 6. An example of estimated distances for different positions (azimuth angles and true distances) is shown in Fig. 7.

A. Signal amplitude and spectral amplitude

For the mean spectral amplitudes we could not observe any systematic change with distance not even in the highest frequency channels (around 10kHz). Nevertheless the spectral amplitudes seem to contain a least some distance information, generating a performance at a low, but above chance level. For the signal amplitude we observed a performance that is better than chance in all aspects even when raw amplitude values are taken. Especially at short distance and regarding near-far confusions performance is very good (less than 3% near-far confusion). This might partially be due to the usage of database sounds which do not represent a natural statistics of amplitudes. On the other hand, the amplitude of a sound at 6 m is only 1/12 of the value at 0.5 m. We observed a small deviation of measured signal amplitudes from theoretical values only for distances of 4m and higher (see Fig. 5). Consequently, we didn't find a substantial difference in performance when using the measured values instead of the theoretical ones. This is probably due to the limited range of distances tested. Correcting for different azimuth angles also had only a small influence on localization performance. If the signal's production amplitude can be predicted, results improve significantly. This means that at least in our test setting familiar distance can indeed be a valuable distance cue. It will depend on the scenario and environment if a prediction of production amplitude is realistic. In unconstrained environments this is surely beyond the limits of current audio processing systems. In more controlled environments (e.g. the system only responds

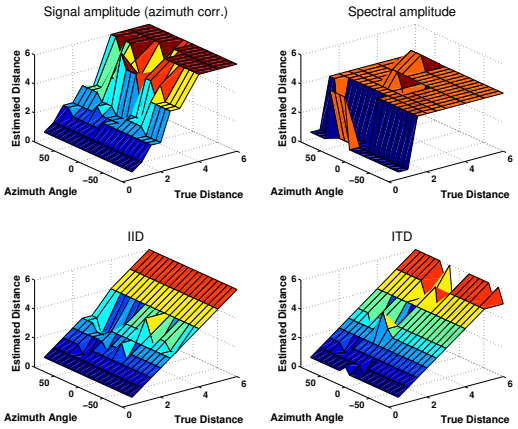

Fig. 7. Estimated distance for four different cues at various true distances and azimuth angles. Data is shown for a single sound.

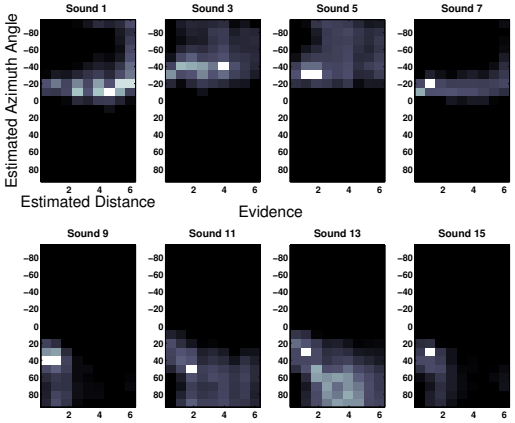

Fig. 8. 2D Position estimation map for a life scenario where a speaker moves in a U-shaped trajectory (see text). Evidence values for different positions are based on combined IID/ITD and signal amplitude similarities. Titles show sound number (see text).

to a small set of a-priori known sounds), though, familiar distance could be an excellent distance cue even in natural environments.

B. Binaural cues

Both IID and ITD show a good performance for distance estimation, with IID being significantly more precise than ITD. As usual for binaural sound localization, it is best to combine IID and ITD measurements. With increasing distance, both binaural cues have a broader distribution of cue values in a frequency channel (see Fig. 3). IID in addition also exhibits a shift towards more extreme values for closer distances (the 'head shadow effect'). On their own IID and ITD can provide an estimation of a sound source's distance with a good precision (e.g. more than 70% correct distance estimation using IID). We observed that the performance of binaural cues depends strongly on the training data. For training we have two sets that were recorded at two different positions of the robot's head in the room (approx. 1 m difference) and different positions of the speaker within the room. Using only one set (Set1) for training and the other (Set 2) for testing the performance decreases substantially: IID mean error increases from 1.14 to 1.8 m and ITD from 1.6 to 1.9 m. This effect is much weaker for the signal amplitude cue: the mean error without any correction factor only increases from 1.8 to 1.9 m. It seems that it is essential to calibrate the system with sounds from many different positions of the source and the robot itself.

To visualize the performance in a scenario with a moving source in the same environment we also recorded test sounds where a freely moving speaker was approaching the robot head from a distance of approx. 6 m and an angle of approximately -30° (sounds 1–5), passing just in front of it (sounds 7–9) and moving away at an angle of approximately 40° (sounds 11–15) on basically a U-shaped trajectory. The localization system was using binaural cues to estimate both azimuth angle and distance plus signal amplitude with azimuth angle correction for pure distance estimation. The results are shown in Fig. 8. Since there is no ground-truth data available, we can't quantitatively evaluate the result. It is obvious that the results are not very precise, but that except for the last sound a coarse 2D position estimation was possible although speaker positions didn't match the position of the speaker in both recordings sets. Also note that there is no tracking or temporal smoothing.

C. Azimuth localization precision over distance

An interesting but so far rarely studied question is the impact of distance on the precision of the azimuth angle estimation. There is a substantial variation in SNR between 0.5 m and 6 m. For an example sound we measured $SNR(0.5m) = 8.7dB$ and $SNR(6m) = 5.2dB$.

The precision of azimuth estimation was investigated in two different scenarios: the first one uses a calibration of the sound localization system for only one reference distance (in our case 1 m) and applies this calibration for all distances. This basically corresponds to a standard approach. The other other method uses the binaural cues to perform a combined estimation of distance and azimuth using a calibration with all distances. The calibration requires more training examples and takes considerably more time but has a chance to incorporate more of the room characteristics.

Our results (see Fig. 9) show that for a calibration with only one distance, performance is best for the calibrated distance and deteriorates with increasing distance (absolute and relative to calibration position). However, using a calibration with all distances the mean azimuth localization error is substantially lower and best performance was measured at the nearest position. We therefore conclude that it is highly beneficial to calibrate the sound localization system explicitly for different distances.

D. Cue integration

An important question is whether the cues we have investigated are complementary or redundant, i.e. if a combination

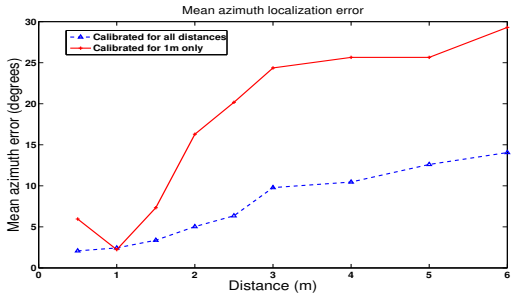

Fig. 9. Mean azimuth localization error for different distances when calibrating only with sounds at 1 m (solid line) and when calibrating separately for all distances (dashed line). The latter approach reduces the localization error by 50% at a distance of 6 m.

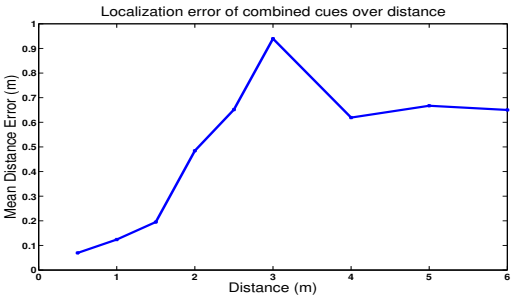

Fig. 10. Mean localization error for the combined cues over distance.

of cues improves the performance or not. We computed cue weights based on the (inverse) measured variances (errors) in the training set. The resulting weights are: 0.12 for signal amplitude, 0.07 for spectral amplitude, 0.6 for IID, and 0.22 for ITD. Note that for the signal amplitude we only employed the azimuth angle correction. The performance of the combined (added) cues is shown in the last row of table I and in Fig. 10. Combined we have a mean distance estimation error of 0.36 m, 27.5% relative error, a correct estimation for 77% and near-far confusion for only 0.12% of all test sounds over all positions. This result is in most aspects better than the IID cue on its own, showing that a combination of localization cues can improve performance.

VI. Summary and Outlook

We have shown a system that uses audio proto objects and biologically inspired localization cues to estimate the distance of a sound source. We got very convincing results in a real-world scenario on a robot head with just two microphones for binaural cues (IID and ITD). For the signal amplitude cue we also demonstrated how an integration of other audio features in the proto object (providing azimuth angle or a hint to production amplitude) could reduce the mean distance error to less than 1 m. In the combination of all cues the mean error was even below 40 cm showing that an extended binaural robot audio system can provide a coarse distance estimation for sounds of different types. We also tested the influence of distance on the precision of horizontal sound localization and found that when calibrating for only one distance, localization precision deteriorates substantially for other distances, while a combined azimuth-distance estimation leads to a generally lower localization error. It should be noted that although we tested a large number of sounds (altogether more than 10000 sounds) in a rather challenging environment, we very likely did not capture the full variability of natural environments. Further tests would be necessary to confirm the results outlined here. The system could be improved further by additional cues with a better cue integration model. If visual input is available, a cross-modal integration will surely improve performance further. Finally, using the audio proto object concept a sound-by-sound tracking of sources over time appears possible.

References

[1] 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.

[2] 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.

[3] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2008)*. IEEE, 2008, pp. 2185–2190.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, (3rd enhanced edition) ed. The MIT press, USA-Cambridge MA, 2001.

[5] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. Int. Conf. Intelligent Robots and Systems (IROS) '05*, Edmonton, Canada, 2005, pp. 509–514.

[6] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 380–385.

[7] K. Nakadai, H. Nakajima, M. Murase, H. Okuno, Y. Hasegawa, and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*. IEEE, 2006.

[8] M. Naguib and H. Wiley, "Estimating the distance to a source of sound: mechanisms and adaptations for long-range communication," *ANIMAL BEHAVIOUR*, vol. 62, pp. 825–837, 2001.

[9] B. S. Nelson and P. K. Stoddard, "Accuracy of auditory distance and azimuth perception by a passerine bird in natural habitat," *Animal Behaviour*, vol. 56, pp. 467–477, 1998.

[10] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," in *Proceedings of the IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS)*. IEEE-RSJ, 2009.

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

[12] J. Weng and K. Y. Guentchev, "Three-dimensional sound localization from a compact noncoplanar array of microphones using tree-based learning," *Journ. of the Acoust. Soc. of America (JASA)*, vol. 110, no. 1, pp. 310–323, 2001.