Spectral cues to source position in robots with arbitrary ear shapes

Tobias Rodemann

2011

Preprint:

This is an accepted article published in Proceedings of ICAR 2011. The final authenticated version is available online at: https://doi.org/[DOI not available]

Spectral cues to source position in robots with arbitrary ear shapes

Tobias Rodemann

Abstract— **The ability to localize a sound source is very important in interaction scenarios where the robot has to face the speaker. It is known that the horizontal position of a sound source can be easily estimated using only two microphones, however, the elevation is more difficult to determine in such a configuration. To deal with these problems the use of special outer ears (so called pinnae) has been proposed in order to allow the use of spectral cues for elevation estimation. Here we compare two algorithms that can extract spectral cues for arbitrary ear shapes and are able to localize a broad class of sounds under challenging real-world conditions. The algorithms run in real-time and are implemented on a real robot-head.**

I. Introduction

With robots entering the domestic and entertainment market, the ability to communicate with humans becomes increasingly important. For the purpose of speech recognition it is generally advisable to know the position of the speaker. This would improve sound source separation [1], facilitate visual speech recognition (e.g. lip reading) and is necessary to robustly identify the speaker in multi-speaker scenarios. It also allows to bring the limited field of view of typical camera setups onto the speaker. Three spherical coordinates have to be extracted: azimuth, elevation, and distance. The first coordinate, the horizontal position, is often considered to be the most important one for interaction scenarios, while the other two coordinates are often ignored. In this article we focus on the second coordinate, the source's elevation. It could provide important information to single out irrelevant sounds (e.g. foot-steps tend to come from below and air-conditioning noise from above) and to determine the height of a speaker (allowing the identification of children).

With a horizontal setup of the microphones the standard binaural cues Interaural Intensity Difference (IID) and Interaural Time Difference (ITD) will not vary much with varying elevation angle, at least in theory. In practice, we have previously shown [2] that binaural cues are indeed capable of providing a combined azimuth and elevation estimation, or a combined azimuth and distance estimation [3]. However, the performance of IID and ITD on their own is insufficient to provide a robust 2D localization under real-world conditions.

Several authors have proposed to use a different type of localization cue to estimate a source's elevation. These approaches use so-called spectral cues that are modifications of the source signal in a position-dependent way. These signal changes are induced by interactions of the sound with the physical structure of the robot. Of very high importance in biology are the outer ears of animals, so called pinnae. It was previously proposed (see e.g. [4]–[6]) to attach structures similar to humans' outer ears next to the robot's microphones to add direction dependent spectral cues that can be used for elevation estimation. Previous approaches had one or more of the following drawbacks: They often used simplified ear shapes, that showed a weak performance with lateral sounds. They also required a substantial amount of human intervention in the selection of the best spectral cues to use. Finally, they only work in a specific frequency range, so that sounds that exhibit no energy in this frequency range can't be localized correctly.

In contrast we present an approach that computes a large collection of audio cues in the full frequency range and learns the relation of these cues to the source's (2D) position. The approach is not principally limited to a specific ear shape (although some ears work better than others) and does not resort to a human designer for hand-picking the best spectral cues. We compare this with a previously proposed approach that also works using arbitrary ear shapes.

Our approach is biologically inspired, mimicking the function of specific neurons in the auditory sub-cortical structures of mammals [7]–[10]. Certain neurons are known to respond selectively to specific spectral features, commonly summarized as 'notch' features although other types of spectral features like peaks are also found. These neurons effectively link a certain spectral feature and the corresponding positions in the outside world for which those features commonly appear. This relation can be learned in a standard training session with a supervision signal (the true position of a sound source) and could also be learned adaptively in an online-system if other cues to position are available.

A. Comparison to Related Work

While a spectral cue-based localization is not commonly used in robot audition, a number of approaches with some similarities to our approach have been presented. Nakashima and Mukao [4] have presented a system with simple, artificial pinnae. Their results were convincing but tests were limited to white noise signals and performance decreased for lateral sounds. As spectral features they concentrated solely on notches of the spectral envelope signal. A similar ear design was employed by Hörnstein et al. on the iCub robot [5]. They used binaural signal differences, but hand-designed the spectral features to rely on, i.e. they decided manually which characteristics in which frequency range are used to estimate the sound's elevation. They showed good results also for speech sounds in a free interaction scenario. Along the same design of ear shapes is the work of Kumon et al. [6] which also employs simple pinnae shapes that produce

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

a direct, simple relation between selected audio features like the position (frequency) of a characteristic notch and the source's elevation.

There are a number of problems associated with these approaches, however. For every ear design the system developer has to manually identify the optimal spectral feature and the parametrization of the function mapping the position of the spectral feature in frequency space to the elevation of the sound source. This function has been approximated by higher-order polynomials [5] or learned from training data [4], but the choice of cues is still manual. Also the simple design of these ears might for some applications not be feasible, e.g. where design constraints are to be considered. There is furthermore the problem that the relevant spectral features often cover only a small subband of frequency space, which means that sounds outside this frequency range are difficult to localize. Finally these approaches employ spectral cues only to estimate the source's elevation, although we could show [2] that they also provide useful information to estimate the source's azimuth position and that this information is at least partially complementary to those extracted from binaural cues.

A more biologically-inspired approach is followed in the work of Neti and Young [9], who are using a neural network to predict a sound source's elevation using data derived from experiments with cats. Gill et al. [11] employed a neural network that uses, instead of binaural spectral signals, several monaural signals from slightly different microphone positions. They argue that animals with movable ears, like cats, could easily vary the ears' position during localization of the sound.

Finally, Saxena and Ng [12] try to predict the spectral features of a sound based on previously learned sound regularities and compare this to the recorded monaural signal. Their microphones are also surrounded by a pinna-like structure.

In [2] we have already demonstrated the usability of spectral cues for sound localization. However in the latter article the focus was on the interplay with binaural cues. In the article at hand we take a closer look at spectral features which are often proposed in biology but rarely used in robotics. We evaluate the performance of different biologically inspired methods and outline how they could be used in future robotic systems.

II. Methods

We are employing a biologically-inspired approach to sound localization and use a human-like robot head with two microphones attached to the sides. The shape of the ears (see Fig. 1) is modeled after a human's ears. The basic preprocessing system is described in more detail in [13], so that only a condensed description will be given in the following.

A. Preprocessing system

We are recording stereo sounds with a sampling rate of 48 kHz. The first processing step is a Gammatone Filter Bank (GFB) [14]–[16] as a model of signal preprocessing in the human cochlea that uses 100 filters with center frequencies between 1–11 kHz. Afterward we compute the signal's envelope through rectification and frequency-specific low pass filtering. To remove stationary background noise, we employ a modified version of spectral subtraction [17], using *Minimum Controlled Recursive Averaging* [18], [19] to estimate the noise level. The final output of the preprocessing is the envelope signal $e_{l,r}(k,f)$ of the left (l) or right (r) microphone at time index (sample) k for the f-th frequency band (GFB channel number).

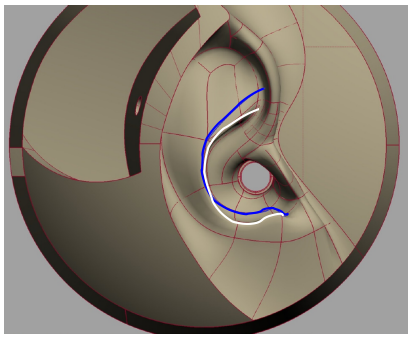

Fig. 1. Shape of (right) pinna employed in our system (CAD data), inspired by human ear shapes. Note that left and right ear shapes are slightly different, as indicated by the blue (left ear) and white (right ear) line outlining the shape of the so-called concha structure of the ears. This asymmetry was used to reinforce binaural differences.

B. Binaural spectral difference vector

The spectral vector $v = \vec{e}_{l,r}(k)$ is a vector of envelope values for all frequencies, i.e. the distribution of signal energies in the different frequency bands. It depends on both the signal's content, e.g. which word is being said, and the Head Related Transfer Function (HRTF), a source position dependent modulation of the signal. While the second part in principle allows us to infer the source's position, the first part is position independent and not principally known. A number of approaches have been put forward to remove the influence of the source signal on the spectral vectors: measuring the sound from two slightly different positions of the head [11] or trying to infer the characteristics of known source signals [12]. In this work we are using a method that was also proposed by [5], that is to subtract the spectral vectors of the left and right ear. Beforehand we apply a log operation on the envelope vector. We call the result the Binaural Spectral Difference Vector BSDV $\vec{d}(k)$:

$$\vec{d}(k) = \log_{10}(\vec{e}_l(k)) - \log_{10}(\vec{e}_r(k)) = \log_{10}\left(\frac{\vec{e}_l(k)}{\vec{e}_r(k)}\right) \quad (1)$$

Due to filtering effects of the robot's head and ears, the BSDV exhibits a number of peaks and notches that are characteristic for different positions (see Fig. 3). BSDVs for different sound source positions are shown in Fig. 2.

C. Segmentation

BSDVs are averaged over different samples with a sample-wise weighting that depends on the total energy of a sample:

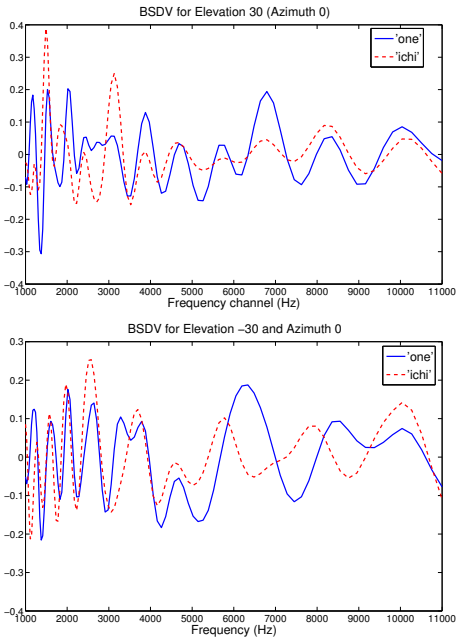

Fig. 2. Binaural spectral difference vectors for two different sounds (English 'One' and Japanese 'ichi' from a different speaker) from two different elevation angles (30 (*top*) and -30 (*bottom*)) at an azimuth angle of 0 (directly in front of robot).

$$\vec{d} = \sum_{k \in segment} A(k)\vec{d}(k) \quad , \tag{2}$$

where $A(k)$ is the signal amplitude in sample k computed as $A(k) = 0.5\sum_f(\vec{e}_l(k,f) + \vec{e}_r(k,f))$, where f sums over all frequency channels. Segmentation is done via a signal amplitude-based threshold operation as specified in [20]. The result is a single BSDV for a single utterance (e.g. a word).

D. Approach 1: Extraction of sparse spectral cues

We observed that even under noise conditions, the position of peaks and notches in the BSDV is rather stable, while the relative height of the peaks varies substantially. We also found that in some cases it wasn't the position of peaks and notches that was invariant but rather the position of the maximum of the slope of the BSDV. We therefore define four classes of spectral cues: peaks (maxima of $\vec{d}$), notches (minima of $\vec{d}$), slope maxima (maxima of $\Delta\vec{d}$), and slope minima (minima of $\Delta\vec{d}$). We found that performance improves when we only use peaks that are above (for maxima) or below (for minima) a certain threshold θ. Before thresholding the BSDV is normalized to a maximum of one. The optimal threshold was found to be at $\theta = 0.3$. The averaged sparseness of spectral cues is 6%, that means only 6% of all possible spectral cues are active for a sound (e.g. on average there are only 6 peaks in a BSDV). Optima are found by looking for frequency channels in the BSDV or its derivative which have higher (or lower) values than both neighboring frequency channels. These spectral cues are shown in Fig. 3. The complete localization approach is shown in Fig. 5 (*bottom left*).

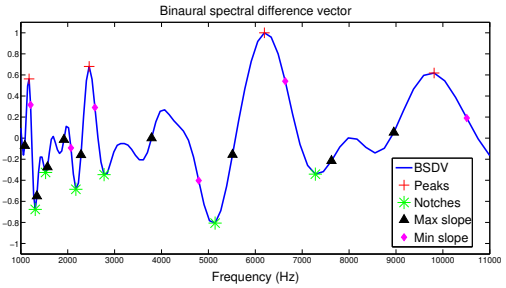

Fig. 3. Binaural spectral difference vector and extracted spectral cues. Consider that only extrema above the threshold $\theta = 0.3$ were used.

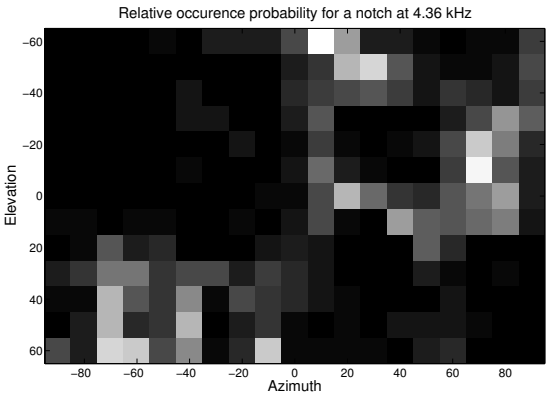

Fig. 4. Probability of occurrence of a spectral cue (here: a notch at 4.36 kHz) as determined from the calibration database. High values are encoded as white, low values as black.

Specific spectral features tend to appear for a number of discontinuous regions in the azimuth-elevation space (see Fig. 4) resulting in high localization errors if the wrong region is chosen. Our algorithm, upon receiving a new sound, looks for spectral cues in the *BSDV*. For every found spectral feature certain regions in azimuth/elevation space are activated, with activity from different spectral cues being additive. Position-dependent spectral cue templates can be built by averaging these template for a larger number of sounds from the same, known position in a calibration sequence. In this calibration process we learn at which sound source positions a certain spectral feature, like a peak in a certain frequency, occurs. For calibration we used a separate training set of sounds over which the average response of spectral filters was computed. For sound source localization, upon detecting specific spectral cues, we invert this relation and activate those regions that produced the detected spectral cues in the calibration set (contributions from different cues are summed). The position that accumulated the maximum evidence was taken as the most likely position of the sound. For a visualization see Fig. 6.

E. Alternative Approach 2: BSDV matching

In a previous paper [2] we have shown a similar algorithm that directly compares BSDVs for source position estimation. For localization we computed the similarity between the currently measured BSDV $\vec{d}_m$ with template BSDVs $\vec{d}(p)$

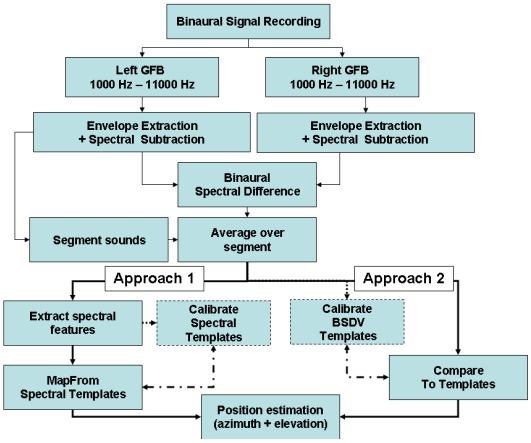

Fig. 5. Diagram showing the two localization approaches. Differences are only in the final processing stages. Approach 1 using sparse spectral features is shown in the *bottom left* while approach 2 (matching of BSDVs) is shown in the *bottom right* .

for every position p (representing azimuth and elevation). Similarity $S(p)$ for a position p is based on the normalized scalar product:

$$S(p) = \frac{\vec{d}_m \cdot \vec{d}(p)}{|\vec{d}_m| \cdot |\vec{d}(p)|} \tag{3}$$

Similarities are directly interpreted as evidence for certain positions analogous to the accumulated evidence for the first approach. The approach is shown in Fig. 5 (*bottom right*).

III. Results

We used a database of 50 sounds from 19 horizontal (between -90 and +90 degrees) and 13 vertical (between -60 and +60 degrees) positions to evaluate our model. Sounds consist of short speech phrases from several human speakers plus some environmental sounds. The recordings were done in a noisy lab of dimensions 12 m x 11 m x 2.8 m with substantial echoes ($T_{60} = 810\ ms$) and background noise (computers, air-conditioning). Sounds were produced by a loudspeaker at a fixed position 1 m away from the robot head. In the recording session the robot head moved to one of the 19*13 = 247 pan/tilt positions (thereby changing the relative position of the loudspeaker). Then the loudspeaker generated all 50 sounds which were recorded and stored to file. Sound preprocessing up to the BSDV was performed in our middleware RTBOS [21]. The final computation and performance analysis was done in Matlab [22]. We also have a full online implementation that can estimate the azimuth and elevation coordinate of a sound source and direct the robot head to gaze at this position. This system employs both binaural localization cues like IID and ITD (see [2], [13]) and spectral cues (approach 2).

For all our tests we have employed 10-fold cross validation to compute performance values. We compute mean azimuth and elevation localization errors (absolute difference between true and estimated azimuth and elevation angle averaged over all sounds in the test database), the hit percentage (how

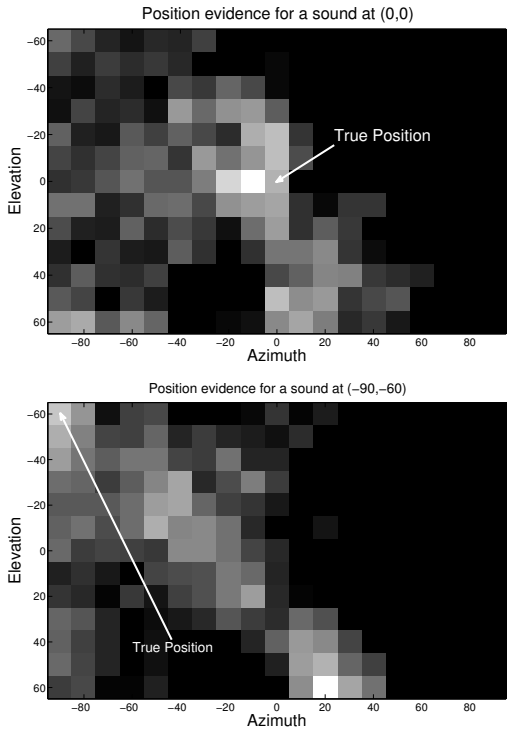

Fig. 6. Accumulated position evidence using the spectral cue approach (1) for two different source positions. In the *top* figure the maximum evidence position is very close to the correct location, but in the *bottom* example the estimation is far off-target. However, substantial position evidence can be found at the correct position (11.9 relative to 15.3 at the maximum).

many sounds were localized correctly, chance level 1/247 = 0.4%), and the relative evidence at the correct position (a^*, e^*), which is computed as $\frac{E(a^*,e^*)-mean(E)}{max(E)}$, where E is the accumulated position evidence in approach 1 or the BSDV similarity in approach 2. The latter measure describes how much evidence was given to the true position relative to the evidence given to all positions and the position which received maximum evidence. This is relevant if spectral cues are combined with other cues (e.g. binaural cues), that provide an independent estimation of the source position.

A. Overall performance

Here we report the results of the proposed algorithms on our test set which are summarized in Table I. The first major result is that both approaches provide a decent amount of information about the position of the sound source regarding it's azimuth and elevation angle. The second approach (BSDV matching) produces substantially better precision for both azimuth and elevation angle estimation. However, the spectral cue approach is able to concentrate more of the position evidence on the correct spot. Both approaches have higher elevation errors than azimuth errors. The hit-rate is 1 in 5 for approach 1 and even 1 in 4 for approach 2 which is far above chance rate.

B. Comparison of single cues

Now we want to analyze the impact of the different spectral cues for the localization performance. In biological

Approach	azi. error	ele. error	correct	pos. evi.
Spectral cues	16.3	18.0	20.0%	0.78
BSDV match	10.7	13.7	26.8%	0.71

TABLE I

LOCALIZATION PERFORMANCE OF THE TWO APPROACHES.

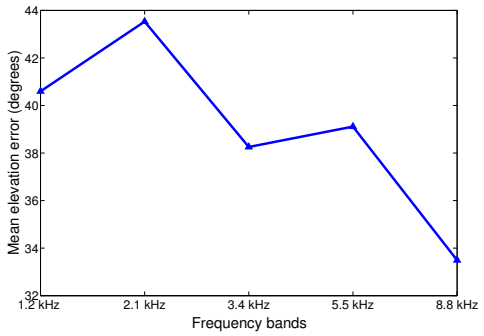

Fig. 7. Mean elevation error for different frequency subbands on their own. Labels on the horizontal axis denote center frequencies for this band.

articles it is often argued (e.g. [7], [9]) that notches (minima in the BSDV) are the main features for spectral localization. We now compare individual contributions of single cue types and also their frequency ranges.

When we look at the contribution of spectral cues from different frequency bands (5 bands with 20 frequency channels each) we observe a trend towards higher performance of the higher frequency bands (see Fig. 7). This is observable for mean elevation error, hit percentage and evidence at correct position. We could not observe a clear trend for the azimuth angle, though.

An analysis of the relative importance of the four basic types of spectral cues (peaks, notches, max and min slope) shows that peaks are on their own the most useful spectral cues (mean elevation error of 25°), while notches are the least useful (mean elevation error of 29°, other cues in between). This is in clear contrast to biological data which suggests that in animals and humans, notches are the prime sources of elevation information. We do not know, whether this is due to the design of our outer ears or due to some difference in the way we compute and process spectral cues.

C. Pure elevation estimation

Hörnstein et al. [5] used binaural cues to estimate a sound's azimuth position, while the elevation was estimated by using spectral cues. Binaural systems can provide a very good azimuth estimation [13], so that we can assume that the azimuth position is roughly given (e.g. via binaural cues). The results for the localization were now re-investigated under the assumption that the azimuth position is already known. In this case the performance of the elevation estimation improves from a mean error of 18.0° without azimuth information down to 12.1° elevation estimation error (and 44% hit rate) when external azimuth information is available (data for approach 1).

Approach	azi. error	ele. error	correct	pos. evi.
Spectral cues	15.0	17.5	31%	0.79
BSDV match	8.4	9.7	42%	0.65

TABLE II

LOCALIZATION PERFORMANCE OF THE TWO APPROACHES FOR THE SECOND HEAD DESIGN.

Fig. 8. Second hardware platform tested.

D. Alternative head/ear hardware

For evaluating if our approach is valid also on other hardware systems we repeated our experiments on a different robot head. We used a human-like mock-up head with (slightly oversized) ears that closely model the human ear (see Fig. 8). Test conditions were similar to the original test, but with a finer (5°) spacing of head bearings. We used identical parameter settings. The results of these tests are shown in Table II. We can see that the performance is similar to the results on the other robot head, but with a higher precision, which might be due to the more elaborate ear shape.

E. Computational effort

The proposed algorithms are computationally very efficient. Starting from a basic binaural sound localization system for azimuth estimation, only a moderate amount of extra computation is required. Therefor our approach is in this regard far more efficient than adding another pair of microphones to determine a source's elevation angle. Of the two approaches presented, the second one (using the BSDVs directly) is conceptually simpler but computationally more costly, due the comparison of the current BSDV with templates for all positions. In the other approach extracting the spectral cues is done quickly and the mapping from those (sparse) cues to position estimates can be done via a simple look-up table. In any way, both approaches easily run in real-time on a single standard desktop computer.

IV. SUMMARY AND CONCLUSION

In this paper we have introduced two approaches for sound source localization that use binaural spectral differences to

determine a sound's azimuth and elevation position. While the localization precision is not very high, it can still provide valuable information about azimuth and elevation. The approaches work using only two microphones and have a very low computational overhead when taking a standard binaural azimuth estimation system as a baseline. We have also shown that both approaches work on two different hardware platforms (with identical parameter settings). In contrast to related approaches we can make use of spectral cues in a larger frequency range, the design of the robot's head and ears is not constrained and no manual intervention is needed to train the system (i.e. the same automated calibration set-up can be used for different robots). If an online training signal is available (e.g. via vision) it would even be possible to perform a continuous adaptation of the system to the current environmental conditions. Our system also provides a combined azimuth and elevation estimation and can therefore be easily integrated with for example binaural localization methods. In a previous paper [2] we have shown that binaural and spectral cues provide at least partially orthogonal information. We have also implemented our approach in an online system that uses both binaural and spectral cues.

In this work we have compared two variations of spectral cues - one that uses the complete binaural spectral difference vector and one that extracts salient features like peaks or notches from this vector. The latter approach is suggested by biological data but turned out to be less efficient than matching the complete BSDV. The probable reason is that using the full BSDV allows the algorithm to exploit more spectral information than just using a sparse set of spectral features. On the other hand, the sparse spectral cues approach was computationally lighter, more sparse in its position estimations, and had a higher relative evidence at the correct position which makes it attractive for combination with other approaches. A potential advantage we didn't investigate is that spectral features are much more localized in frequency so that concurrent, partially overlapping sounds might have a less disturbing impact.

We also found that at least for our human-inspired ear designs the relevant spectral cues were not limited to spectral notches but all four types of cues contributed substantially. In accordance with biological data we saw a larger impact of higher frequencies.

The algorithm would probably be improved by optimizing the range of frequencies that are used in the GFB, e.g. having more frequency channels for the higher frequency bands.

Along a different line of research it is conceivable to optimize the shape of ears to provide the best possible spectral cues. This process can for example be based on evolutionary design optimization.

REFERENCES

[1] 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.

[2] 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.

[3] T. Rodemann, "A study on distance estimation in binaural sound localization," in *Proceedings of the IEEE/RSJ conference on Intelligent Robots and Systems (IROS)*, 2010.

[4] H. Nakashima and T. Mukai, "3D sound source localization system based on learning of binaural hearing," in *IEEE International Conference on Systems, Man and Cybernetics*. IEEE SMC'05, October 2005.

[5] 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.

[6] 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.

[7] K. A. Davis, R. Ramachandran, and B. J. May, "Auditory processing of spectral cues for sound localization in the inferior colliculus," *JARO*, vol. 04, pp. 148–163, 2003.

[8] A. J. King, J. W. H. Schnupp, and T. P. Doubell, "The shape of ears to come: dynamic coding of auditory space," *TRENDS in Cognitive Sciences*, vol. 5, no. 6, pp. 261–270, 2001.

[9] C. Neti and E. D.Young, "Neural network models of sound localization based on directional filtering by the pinna," *J. Acoust. Soc. Am*, vol. 92, no. 6, pp. 3140–3156, 1992.

[10] 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.

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

[12] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2009, pp. 1737 – 1742.

[13] 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, pp. 368–373.

[14] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, *Auditory Physiology and Perception*. Exford: Pergamon, 1992, ch. Complex sounds and auditory images, pp. 429–446.

[15] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, pp. 103–108, 1990.

[16] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank,," Apple Computer Co., Technical Report 35, 1993.

[17] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[18] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 446–475, September 2003.

[19] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[20] 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.

[21] 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.

[22] *Matlab*, 14th ed., The MathWorks, www.mathworks.com.