Audio Proto Objects for Improved Sound Localization

Tobias Rodemann, Frank Joublin, Christian Goerick

2009

Accepted for IROS 2009, October 11 – 15, 2009 in St. Louis, USA

Audio Proto Objects for Improved Sound Localization

Tobias Rodemann, Frank Joublin, and Christian Goerick

***Abstract*— In this article we present a new framework for auditory processing that combines feature extraction and grouping processes to form what we call audio proto objects. These proto objects combine an arbitrary number of audio features in a compact representation that allows a more precise sound localization and also better interfacing to behavior-control in robotics. We compare our standard sound localization system with the new approach in several scenarios to demonstrate the potential of the new approach.**

I. Introduction

When talking about robot audition two specific sub-tasks dominate in literature: Speech recognition including sound source separation to improve the signal-to-noise ratio on the one side and sound localization on the other side. Both processes transform a low-level audio signal into a high-level, more symbolic representation for generation of behavior. While this process is well-defined in speech recognition, for the task of localization the transition from the signal to the behavior level is often designed ad-hoc. In most applications ([1]–[3]) low level localization features like ITD and IID are converted into probabilities for different positions. In a second stage the currently most likely position of the sound source is extracted, e.g. by finding the peak in a position map. In the final stage the robot's attention or gaze is shifted towards this position. Because one normally wants to avoid responding to spurious background activity, a threshold operation is often applied. Furthermore, since instantaneous single-sample measurements are unreliable under real-world conditions, measurements are normally integrated over time to smoothen the result. Finding an optimal integration time constant under varying conditions is difficult. It is also challenging to decide when to read-out the position estimation. Earlier position estimations are often too noisy since they use only part of the available cues while later responses (e.g. second onset in same word) are often affected by echoes to some degree. A different type of problem is that it is difficult to base the decision whether to attend to a stimulus or not purely on the signal's position or energy. Other audio features like sound length or pitch might be more suited to separate relevant from 'noise' stimuli.

To solve these issues we propose to use a concept that was inspired by the work of Bregman [4] on auditory scene analysis (ASA). Our idea is to perform bottom-up segmentation of audio signals along the time or frequency domain and then compute compressed audio features over the full segment length. These compressed features, plus time information on the segment, are combined into what we call an audio proto object. These audio proto objects may correspond to whole utterances or tones, but might as well be combinations or fragments of sound objects. We will show that this representation is well suited for tasks like sound localization with selective attention. Both precision and reliability of sound localization are improved and it is easier to filter irrelevant stimuli before moving the robot. In this article we will first describe how we perform the basic segmentation process based on the signal energy only. After the segmentation process we outline how we represent feature values over the full segment using averages or histograms. We will then describe in more detail the audio features used in our system: signal energy, sound direction (azimuth) via IID and ITD, energy slope and the length of segments. Other features like pitch and spectral energy have been tested, but will not be discussed in the scope of this paper. Later on we will report on several experiments in a real-world environment we have done to compare a standard sound localization approach as described in [1] with our new concept. We demonstrate that the mean localization error is comparable to standard approaches with well-tuned integration constants and in some scenarios even better. We also show that in scenarios with several alternating speakers the integration over different utterances of the same speaker is possible. Finally we show an example where, based on two simple filters, the majority of background noise signals could be ignored in an experiment with a humanoid robot in free interaction. Figure 1 shows the system's architecture with feature preprocessing, segmentation, audio proto object generation, filtering, grouping, and motor control modules.

A. Comparison to related work

The term audio proto objects is closely related to both audio streams ([4], [5]) and visual proto objects [6]. Audio streams are the result of a segmentation process operating on a number of audio features. Our approach uses a strongly compressed representation of audio features in the form of mean values or histograms. While audio streams are better suited for speech recognition, they are often too cumbersome to be a basis for operations on the behavior level. We argue that for many robotics applications, there is a number of problems which are not tackled by the current approaches. One is that the robot needs to distinguish between relevant stimuli like user commands and distractors such as phone ringing, foot-steps, or people talking with each other. However, this separation has to be flexible depending on the situation. We also believe that robots need to understand how many sound sources are around them, which characteristics they

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

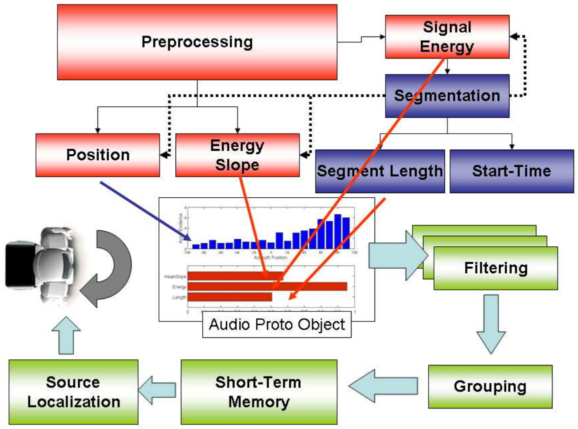

Fig. 1. System architecture (the preprocessing module is described in more detail in Fig. 3).

have and how sound sources are related to each other (i.e. who is talking to whom). These are difficult challenges that will require a higher level representation of audio signals. Due to the length of raw audio signals in complex situations (imagine following a dialog) these representations will need to be very condensed, otherwise the relevant characteristics can't be extracted.

II. Audio proto objects

In this section we introduce the concept of audio proto objects as a high-level, compact representation of audio signals for linking with other sensory modalities or behavior control in robots. We assume that, after sound acquisition and preprocessing, a number of audio features are computed. One or more of these features are used for the segmentation process that defines the borders of a segment. The segmentation is described below. The next processing stage computes compressed audio features over the whole segment and also calculates derived features (start and length of the segment) based on the segmentation process. Finally, compressed audio cues and derived features are combined to one entity that we term audio proto object.

A. *Segmentation process*

One of the most critical aspects for the generation of audio proto objects is the definition of segments. In this work, we use only a simple energy-based segmentation process. We assume that relevant sounds are sequential, so that a separation in time is sufficient. A proto object starts when the signal energy exceeds a threshold θ and ends when the energy falls below this threshold. The parameter θ depends on the hardware characteristics and needs to be adapted to the background noise level. Fig. 2 gives an example of the segmentation process.

Our approach is currently limited to situations where speakers alternate without any overlap. Nevertheless a number of realistic scenarios will be of the type that can be handled in our approach and literature has shown a number of solutions for separating concurrent sounds in real-world applications [5], [7], [8].

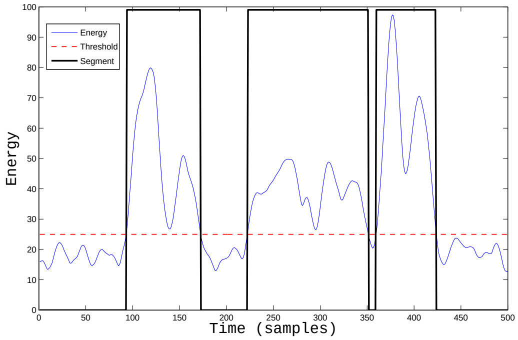

Fig. 2. Example of energy-based segmentation for sound data recorded on our robot Asimo. The energy was computed after the spectral subtraction removed stationary background noise.

B. *Feature compression*

The feature compression stage integrates cues over all samples and provides a description of the feature over the full segment length. The new representation can be a scalar value, like average signal energy, or a vector over different frequency channels or positions. In any case, the representation is independent of the size of the segment. Specifically, in the audio proto object, energy is represented as the mean value over all samples in the segment (of length L).

$$P_{energy} = \frac{1}{L}\sum_{s\in S} A(s) \quad , \tag{1}$$

where $A(s)$ is the sum of signal envelope values over all frequency channels in sample s and S the set of all samples in the segment. The representation of the localization is the accumulated position evidence for all samples:

$$P_{position}(\alpha) = \sum_{s\in S} E(\alpha, s), \tag{2}$$

where α is the azimuth angle of the source, and $E(\alpha,s)$ the evidence for azimuth angle α in sample s. $E(\alpha,s)$ is integrated over time with a constant τ, see [1].

III. System

The basic system architecture (see Fig. 1) is based on the one presented in [1], extended by several preprocessing elements (see Fig. 3 (left)) and modules for the generation, filtering, and grouping of proto objects. Localization as the main audio feature used in this article is based on the Interaural Intensity (IID) and the Interaural Time Difference (ITD) as cues. A model of the precedence effect is used to reduce the impact of echoes and spectral subtraction is employed to reduce background noise. Sound data was recorded on a humanoid robot head modeled after Honda's Asimo, see Fig. 3 (right). We are using two human-inspired ears mounted on the sides of the robot. The head is in a noisy, very echoic ($T_{60} = 810ms$) lab room of size 12 x 11 x

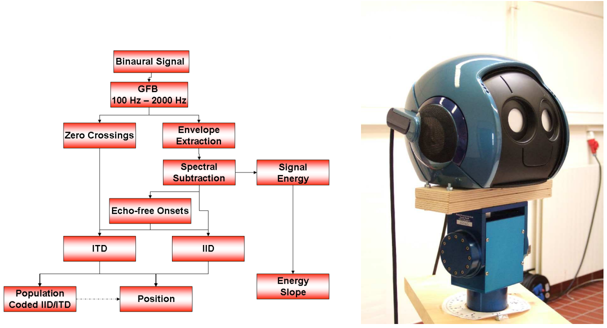

Fig. 3. *Left*: Sketch of system's preprocessing architecture. *Right*: Asimo-like robot head with two human-inspired ears mounted on a pan-tilt element.

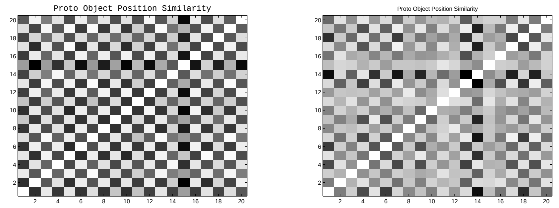

Fig. 4. Scalar product similarity measure between different proto object positions for two different speakers at azimuth angle 10° and -10° (*left*) and at 30° and 60° (*right*). The dialog structure is visible in both plots.

2.8 m. We are using a set of 57 different sound files recorded at different positions in this room. A subset was used for calibrating the audio-motor mapping, the remaining 47 files were used for evaluation. We used a Gammatone Filterbank (GFB) [9] with 100 frequency channels that spans the range of 100 - 2000 Hz, where, due to background noise in this frequency range, performance for our standard system still shows potential for improvement.

A. Filtering, grouping and short-term memory

A big advantage of the proto object concept is that proto objects can be easily filtered depending on their condensed features. As an example, proto objects that are too short or don't have enough energy, can be neglected for sound localization or other behaviors. The proto object concept can be extended to group proto objects from the same source together since they have similar features. For grouped proto objects, feature measurements can be integrated thereby improving localization performance. While the standard temporal integration approach makes some (implicit) assumption about the auditory scene (e.g. sounds close in time are from the same source) the proto object approach allows a more explicit and flexible integration.

We use position to group audio proto objects together, but in certain situations also cues like pitch or spectral content might be helpful to distinguish different sound sources. In scenario 2 (see below) a grouping of proto objects could even be done based on energy and segment length (due to slightly different characteristics of speakers' speech volume and segment length), but this probably would not apply in most scenarios.

We do not perform an offline clustering of proto objects but rather employ a sequential procedure where for every new proto object it is decided if it is integrated with other proto objects. Grouping is done by evaluating the position vector $P^{new}_{Position}(\alpha)$ of the new proto object and comparing it to the position vectors $P^{j}_{Position}(\alpha)$ of all proto objects in short-term memory. If the position vectors are similar ($S(new, j_{min}) > T_S$), the closest proto object j_{min} is updated, otherwise a new entry in memory is created. An update is done by adding the position evidence from the new proto object to the old representation:

$$P^{j_{min}}_{Position}(\alpha) \rightarrow P^{j_{min}}_{Position}(\alpha) + P^{new}_{Energy} * P^{new}_{Position}(\alpha) \qquad (3)$$

All position evidence is multiplied with the proto object energy P^{new}_{Energy}. Additionally, activity decays exponentially over time and very weak proto objects are removed from short-term memory. Similarity $S(i, j)$ is based on the scalar product of normalized position evidence vectors of the two proto objects. Fig. 4 shows the pairwise position similarity of proto objects for two dialog settings. Based on this figure we estimated the optimal similarity threshold to be $T_S = 0.6$.

B. Localization with population-coded cues

Our standard sound localization system maps pairs of ITD and IID values to azimuth position candidates using a pre-calibrated audio-motor map. Cue pairs are measured for specific frequency channels whenever an onset occurs. There are normally only very few onsets in a single word for a specific frequency channel. Considering that cue measurements are noisy and, especially for ITD, ambiguous, mapping single cue pairs naturally produces a large number of candidate positions. Only in the integration over many channels and over time the necessary robustness and precision is gained. This integration is a summation of position estimations over all channels and a leaky integration over time. Because the audio proto object concept allows a grouping of different frequency channels and onsets over time, a better localization performance should be reachable. To test this hypothesis, we combined cue measurements over the complete segment of the audio proto object and then mapped the result to a position estimation. In order to retain information about the distribution of cue values in the proto object, individual measurements are re-encoded into a population code. For ITD and IID each, we use a set of nodes with response centers at -0.9, -0.8,...,0,...,0.8,0.9. Every single measurement of ITD or IID leads to an activation in the nearest nodes (with a Gaussian distance kernel of width 0.1) and all measurements for a single audio proto object are added up. There is one population code vector for each frequency channel so that the population-coded cue representations for a single proto object have a size of 2 (ITD+IID) * $N_{FreqChannels}$ * N_{nodes}. During calibration we measured the population response by averaging over all 10 calibration files. Cue to position mapping is performed by combining all cue measurements for the whole proto object as described above

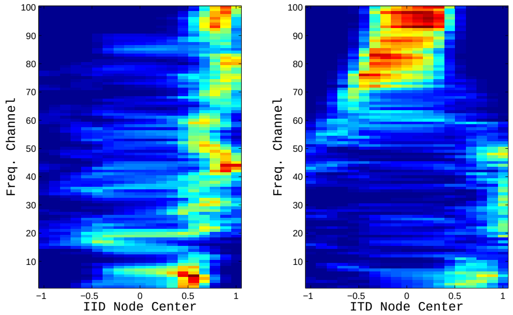

Fig. 5. Population coded representation of binaural localization cues (*left*: IID , *right*: ITD) for azimuth α = 50°. Data is integrated over 10 different sounds for each angle.

and then comparing the population response with the stored population responses for all positions. The evidence $W(\alpha)$ for a specific angle α is computed as the scalar product over nodes between the measured values $m_{m,c}(n)$ and stored representation $M^{\alpha}_{m,c}(n)$ for frequency channel c, cue m (1 = IID, 2 = ITD), and population node n, summed over all frequency channels and cues in the population:

$$W(\alpha) = \sum_{m=1}^{2} \sum_{c=1}^{N_{Freq}} < M^{\alpha}_{m,c}(n), m_{c,m}(n) > \qquad (4)$$

All vectors in the scalar product are beforehand normalized to mean 0 and norm 1. Fig. 5 gives an example population code representation of localization cues in the proto object for a sound at 50° azimuth.

IV. Results

For the analysis of the proto object-based localization we recorded a number (47) of sound files from different positions and compared localization results for varying parameter settings. In a first scenario, source positions vary slowly from 90 to -90 degrees azimuth bearing, and all sounds are played in sequence for each position. This corresponds to a sound source that slowly moves from right to left. In such a scenario a temporal integration of localization cues is very beneficial. The second scenario simulates a dialog situation where two sound sources (speakers) talk in alternation. Both speakers count from one to ten (in different languages). The two scenario situations are sketched in Fig. 6.

In the following, errors are given as mean azimuth error and the percentage of correct estimations (i.e. error = 0). The peak position in the localization map (over the whole sound file for the standard approach) is taken as the estimated position. This procedure favors the conventional approach since it implicitly solves the problem of determining the optimal point for the read-out of the position estimation.

A. *Temporal integration vs. proto object concept*

Comparing the performance of different sound localization approaches is difficult since the results depend on a large number of factors (room conditions, recording hardware, test sounds, and others). We therefore compare our previous sound localization system as outlined in [1], which showed very good performance under high-noise conditions, with our new, proto-object-based approach, that works on top of the old system. We vary the temporal integration constant τ from 10 ms to 50s. Our results (Table I) for scenario 1 show that the performance of the standard system increases with larger values of τ, mostly due to the integration over different sounds.

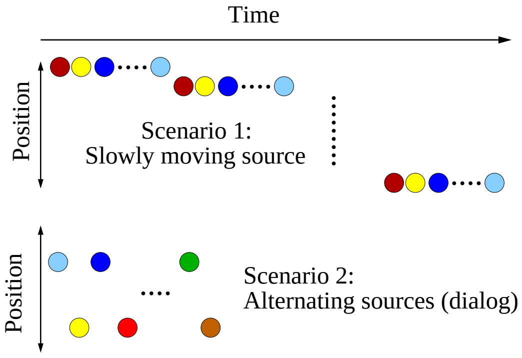

Fig. 6. Sketch of the two main scenarios used in this paper. Spheres denote different sounds.

τ	mean loc. error	percent correct
τ=10 ms	8.0°	46.8%
τ=100 ms	7.8°	47.7%
τ=1000 ms (*single*)	7.3°	50.1%
τ=1000 ms	6.5°	51.9%
τ=5 s	4.7°	56.6%
τ=20 s	3.9°	62.6%
τ=50 s	3.9°	63.3%

TABLE I

Azimuth error for scenario 1 with our standard system and different values of τ . In the *single* setting, integration does not extend over more than one sound file each.

If we apply the same strategy for the alternating sound scenario (2) the results are different (Table II). We used three different settings, one with the sound sources at 10° and -10° (setting 1), one with sources at 40° and -40° (setting 2) and finally 30° and 60° (setting 3). The last setting is especially challenging because both sources are on the same side of the head and therefore mislocalization can arise easily. In these settings temporal integration does not make sense for more than a few 100 *ms*.

The problem in the dialog scenario is that with increasing time constant past measurements get an increasing influence on the current position estimation. As a result the localization will either localize only one of the two sources (as for setting 2) or average over the two positions (as in setting 1 and 3). These results demonstrate that for situations which resemble dialog scenarios, time constants of integration would optimally be in the range of 100 - 1000 *ms*. However, our results in scenario 1 have shown, that, if there is just one source, localization precision could be improved substantially when using longer integration constants.

In comparison, the audio proto object approach uses

τ \setting	Setting 1 (10/-10)	Setting 2 (40/-40)	Setting 3 (30/60)
τ=10 ms	3.5° (65%)	10.5° (35%)	7.5° (35%)
τ=100 ms	3.0° (70%)	9.0° (30%)	8° (30%)
τ=1000 ms	3.0° (70%)	10.0° (25%)	9.5° (30%)
τ=5 s	6.0° (60%)	20.5° (35%)	12° (25%)
τ=20 s	8.5° (55%)	37.5° (45%)	13° (15%)

TABLE II

Azimuth error for the standard system with sources at different azimuth angle settings

Filter setting	Proto object approach	+ population coding
ALL	12.4°	8.3°
Top99	12.2°	8.0°
Top90	10.6°	6.7°
Top80	9.0°	5.5°
Top60	6.7°	4.3°
Top40	6.3°	3.9°
Top20	6.1°	3.9°

TABLE III

Performance of proto object system for scenario 1 with different settings of energy filtering . The rightmost column depicts results for the proto object approach with population coding of cues.

Setting	Best standard	Proto Object	+ pop. code
Sources at 10/-10	3.0° (70%)	3.5° (65%)	1.5° (85%)
Sources at 40/-40	9.0° (30%)	8.5° (35%)	6.0° (50%)
Sources at 30/60	7.5° (35%)	8.0° (35%)	4.5° (55%)

TABLE IV

Comparison of azimuth localization precision for standard and proto object approach for variants of scenario 2.

Scenario	Estimated Positions
Two sources (-10/10)	-10 / 20
Two sources (-20/20)	-30 / 30
Two sources (-30/30)	-30 / 40
Two sources (-40/40)	-40 / 50
Two sources (-50/50)	-50 / 60
Two sources (-60/60)	-70 / 70
Two sources (-70/70)	-70 / 70
Two sources (-80/80)	-80 / 90
Two sources (-90/90)	-80 / 90
Two sources (30/60)	30 / 70
Four sources (-20/40/-50/70)	-30/50/-50/80

TABLE V

True and estimated positions using the proto object approach and grouping based on position similarity after ten utterances each.

only an integration over samples that are grouped by the segmentation process. Directly comparing the performance of standard and proto-object approach is difficult because, due to the segmentation process, there are on average 1.5 proto objects generated for each sound played. Some of these proto objects have a very low energy or length and also bad localization performance. We therefore decided to filter proto objects with a low energy and use only the remaining ones for measuring the localization performance. Table III gives result for different settings of the filter, where *ALL* means that no filtering is used and *TopXX* that only the top *XX*% proto objects (in terms of their energy) are evaluated. The right column provides the results for the population coded sound localization (see section III-B).

Performance increases substantially when working only with louder proto objects. Using only 60% of the proto objects (roughly one per sound as for the standard approach) the results are better than the ones for the *single* setting in the standard approach where information can't be integrated over several sounds. It is also noteworthy that the population coded approach is substantially better than the single cue mapping (a reduction of more than 30% in localization error).

For the second scenario (alternating speakers) the results are shown in table IV. Again, low-energy proto objects have been filtered out (1 or 2 per setting). The results are comparable to the performance of the standard approach for a setting of the integration constant that is well adapted to the timescale of the dialog. With the proto objects plus population coding results are even substantially better. At this point we are not even using the full potential of the proto object approach. In the next paragraph we will show how the grouping of proto objects according to their features improves the localization further.

B. Grouping of proto objects via position

We used the position-based grouping process to improve localization precision and to estimate the number of sound sources in the different scenarios. For different settings of the dialog scenario two main sound sources emerge at the positions shown in Table V. In addition, one or two additional weaker sources were found due to localization outliers but they have very low accumulated position evidence and will disappear over time.

The results show that the two sources can be extracted and their position localized with an average error of 5.5°. Even in a scenario with four separate speech sources the grouping process correctly determined the number of sources and their positions. Is has to be noted, that with increasing number of sources the grouping process becomes more and more difficult. While two sources can be correctly identified with a broad range of values of the similarity threshold, more tuning is necessary for more sources. This process can be improved if more separating cues are available and similarity is based on several cues.

When applying the same approach for the first scenario, the system can for most of the time follow the current source position (i.e. just one proto object with correct position estimation), being able to ignore rare spurious proto objects with wrong position estimations. The system only fails when the source switches sides since at this point the distribution of position evidence values changes substantially. Quickly, however, a new proto object at the correct position will emerge. When taking as position estimation the peak position from the strongest proto object, the mean localization error

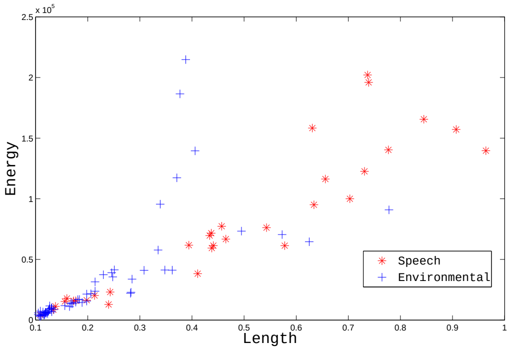

Fig. 7. Length and energy of audio proto objects for two types of sounds: environmental noise and robot directed speech.

is 4.0°. With the population coding approach the precision is very high (2.4° average error and 77% correct) and the target is never lost. Note that this process is iterative and based only on the measured position. Most of the time the system can correctly estimate the number and position of sound sources.

C. Filtering of environmental sounds

Another application of the audio proto object concept is the filtering of environmental sounds. We recorded sounds on our humanoid robot Asimo while it was powered up. Sounds are of two types: environmental sounds like mouse-clicks, footsteps or door slamming, which the robot is supposed to ignore, and speech directed to the robot, to which the robot should orient to. Audio proto objects were extracted for both databases and feature values for the two sound categories compared. The databases contained 55 environmental sounds and 25 speech commands recorded in a realistic scenario.

It turns out that most environmental sounds are rather short (mean 0.36 s compared to 0.8 s for directed speech) and have a low mean signal energy (mean 52880 compared to 97000 for robot directed speech). In Fig. 7 proto object length and energy are plotted. Using a simple threshold on length (0.5 ms) and mean signal energy (30000), 80% of the environmental sound proto objects can be filtered out while 92% of the speech signals can pass through. We have successfully implemented the environmental sound filtering mechanism on our Asimo robot as part of a larger integrated system similar to [10]. As a result of the filtering operation, the robot almost exclusively responds to humans calling the robot, ignoring most of the background noise. This was reached without any speech-specific audio features. However, integrating more cues like pitch or formants is straightforward and could enhance performance.

V. Summary and Outlook

In the spirit of Bregman's Auditory Scene Analysis we have introduced a new concept for sound processing in robotics which consists of an energy-based segmentation process and a feature compression and concatenation stage. The resulting audio proto objects are a framework for increasing sound localization performance in typical robotic scenarios, including a higher precision in multi-source scenarios, integration over several utterances of a speaker, combination of different cues for grouping processes over time, and filtering of specific sounds. We have also shown that the population coding of localization cues for the entire proto object can furthermore reduce the localization error by about 30%.

The audio proto object concept should be extended by improving the segmentation process, for example through source separation. This would extend the concept of segment into the spectral dimension and allow a treatment of concurrently active sources. Other necessary extensions are more audio cues for segmentation and grouping, like spectral structure, formants, or HIST features [11]. These additional features would allow us to extend the proto object concept to more types of scenarios. We also plan to extend the proto object based localization to 2D, combining binaural and spectral cues as described in [12]. Audio proto objects can be combined with visual proto objects [10] since their structure is similar. Both types of proto objects have a specific position in time and contain compressed feature representations. Combining the two modalities on the proto object level could lead to some interesting applications.

References

[1] 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.

[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] H.-D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers," in *Proc. Int. Conf. Intelligent Robots and Systems (IROS)*, Nice, France, 2008.

[4] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[5] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis*. IEEE Press, 2006.

[6] B. Bolder, M. Dunn, M. Gienger, H. Janssen, H. Sugiura, and C. Goerick, "Visually guided whole body interaction," in *IEEE International Conference on Robotics and Automation (ICRA 2007)*. IEEE, 2007.

[7] 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.

[8] M. Heckmann, F. Joublin, and E. Körner, "Sound source separation for a robot based on pitch," in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*. IEEE, 2005, pp. 203–208.

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

[10] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmüdderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, 2008.

[11] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A closer look on hierarchical spectro-temporal features (HIST)," in *Proc. INTERSPEECH 2008*. Brisbane, Australia: ISCA, 2008.

[12] 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.