Filtering environmental sounds using basic audio cues in robot audition

Tobias Rodemann, Frank Joublin, Christian Goerick

2009

Accepted for ICAR 2009, June 22 – 26, 2009 in Munich, Germany

Filtering environmental sounds using basic audio cues in robot audition

Tobias Rodemann, Frank Joublin, and Christian Goerick

***Abstract*— In this article we present an approach for separating robot-directed speech from environmental sounds for applications in robot audition under high noise conditions. We introduce a new framework for audio processing that combines feature extraction and a grouping process 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 filtering of environmental sounds using relatively simple audio cues like signal energy or segment length. We demonstrate that our system can be a first step towards a selective and adaptive auditory attention in real-world robotics scenarios.**

I. Introduction

The two main applications in robot audition are probably sound localization and speech recognition. Especially for the latter task the separation of relevant (speech) signals from background noise, environmental sounds or concurrent speech activity has always been of high concern due to the strong deterioration of performance with increasing noise level. While the removal of stationary background noise is relatively straightforward, dealing with concurrent sounds is very challenging (see [1] for a review). State of the art methods are capable of separating several concurrently active speakers and provide a sufficiently clear signal to speech recognition [2].

A problem that is often ignored in this context is the question to which sounds the robot should attend. This is very important for speech recognition but also for sound localization. In free audio interaction, i.e. using the robot's microphones instead of a close-talk microphone, all sounds generated in the environment compete for the robot's attention. A sound localization system can even respond to whispering or mouse-clicks [3], so that the robot can be easily distracted. Applying a threshold on the signal energy is a common remedy but this acts on a very low sensory level and is not effective for louder distractor sounds like impact noise. Separating speech from non-speech would also not be sufficient and is furthermore difficult to achieve under real-world conditions on a robot.

The specific scenario we target in this article is a humanoid robot (ASIMO) that uses two human-inspired ears mounted on the sides of the head to localize sound sources and orient its gaze towards them. After turning its head the robot can start interacting with a human partner. Due to the limited field of view of the robot's cameras it is essential to direct the robot's attention to the right position. Since the ears are close to the robot's fans and motors and we operate in a typical noisy, echoic robot lab, sounds can easily distract the robot and pull its attention away from the interactor. The sound processing system described in this work is embedded into a larger, interacting system similar to the one outlined in [4].

We present a framework for transforming audio signals into a compact representation called audio proto objects. On this representation level, behavior selection (e.g. to listen or not to listen), can operate in a simple and flexible way. Audio proto objects combine compressed representations of basic or higher-level audio features like signal energy, source position, or spectral energy for an audio segment.

Our system targets to filter environmental sounds that are irrelevant for the robot's behavior. Typical examples are footsteps, mouse-clicks, door slamming, or people talking in the background. In contrast, the system should respond to (by orienting its gaze towards) calls from people directed at the robot. We want to avoid using specially designed audio features but rather use simple, basic elements of audio processing. We note explicitly that the target is not just to separate speech from non-speech sounds since one of the biggest classes of distractors is people talking with each other but not towards the robot. The solution we aim at should be flexible and on a higher level than standard sound processing systems. Certain types of mid-level audio features like pitch or formants were not used in the analysis, because, although potentially interesting, they perform badly in scenarios with a very high noise background. The purpose of the application described in this article can be summarized as turning raw audio data into proto objects and then categorizing them to be either background signals or robot directed calls. While the former class is to be ignored the latter type of sounds should trigger a gaze change of the robot toward the estimated position of the proto object.

Figure 1 shows the system's architecture with feature preprocessing, segmentation, audio proto object generation, similarity computation, filtering, and gaze change modules. In the next sections we will first describe the basic system architecture and the audio features that are computed and then proceed with a more detailed explanation of the audio proto object concept, which is the main contribution of this paper. We will then analyze the performance of our concept for sound localization and investigate the potential of different audio cues for differentiating between the two relevant sound classes (background and calls).

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

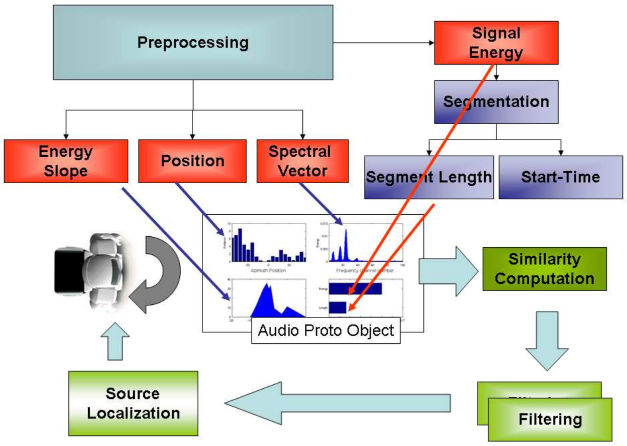

Fig. 1. System architecture (the preprocessing module is described in more detail in Fig. 2).

A. Comparison to Related Work

Sound processing often has a strong focus on two types of sounds - music and speech. Other types of sounds, although dominating in natural environments, are rarely investigated. There have been a few attempts to separate speech from non-speech signals [5] and to categorize environmental sounds [6], [7]. Early work of Gaver [8], [9] has even hinted that substantial behaviorally relevant information can be extracted from environmental sounds. All in all, however, research has been very limited compared to the large body of speech related work. In the field of robot audition, where environmental sounds occur abundantly, research on how to deal with non-speech sounds has been very sparse.

The term audio proto objects is closely related to both audio streams [1], [10] and visual proto objects [11]. Audio streams are segments in time and frequency space and well suited for speech recognition. But for behavior selection on a robot, they are often too cumbersome since they represent audio data as a raw sequence of samples without a temporal grouping. The result is a high-dimensional representation (typically 100s of samples per second are used) with significant noise in individual samples and substantial variations of audio features on a sample-by-sample basis. We propose to use a more flexible and light-weight representation of audio signals that facilitates an adaptable categorization and filtering of sounds based on a number of audio cues.

II. System Cue Processing Architecture

The basic system architecture is based on the one presented in [3] extended by several preprocessing elements (see Fig. 2) and modules for the generation, comparison, and filtering of proto objects. Sound source localization for the azimuth position 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.

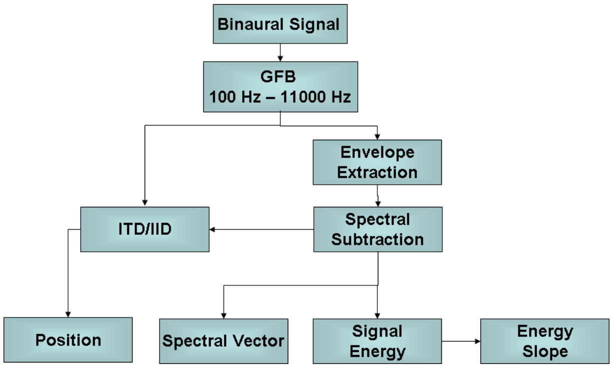

Fig. 2. Sketch of the system's preprocessing architecture.

Fig. 3. Our humanoid robot ASIMO with custom-made, human-inspired ears.

In contrast to the work described in [3], [12], where all IID/ITD pairs are mapped individually to position candidates, our proto object-based localization system collects cue measurements for the full proto object (i.e. over all frequency channels and samples) and maps them as a whole toward candidate positions. The mapping is learned in a calibration with 10 sounds per position.

Sound data was recorded on Honda's ASIMO robot using two human-inspired ears mounted on the sides of the robot (Fig. 3). The robot is in a noisy, very echoic (T_{60} = 810 ms) lab room of size 12 x 11 x 2.8m. We used a Gammatone Filterbank (GFB) [13] with 100 frequency channels that span the range of 100 - 11000 Hz.

A. Signal energy, energy slope, and spectral vector

After applying the Gammatone filterbank and extracting the envelope signal for every frequency channel we use a spectral subtraction (as detailed in [3]) to remove the stationary background noise. The resulting signal at time (sample) s and frequency channel c we term $A(s,c)$. The signal is low-pass filtered and sub-sampled to 1000 Hz, down from 48000 Hz sampling rate at the microphones. The signal energy $A(s)$ is then derived as the sum over envelope signals in all channels:

$$A(s) = \sum_{c} A(s,c) \quad . \qquad (1)$$

The signal energy will be used to define segment borders as described below. A derived feature is the energy slope, the difference in signal energy between two consecutive samples:

$$\delta A(s) = A(s) - A(s-1) \quad . \tag{2}$$

This is a simple measure of amplitude modulation in the signal. The spectral energy vector $\vec{A}(s)$ contains the distribution of energies over all frequencies for a specific sample s.

III. 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 sounds 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 is used for the segmentation process that defines the borders of a segment. 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 important aspect 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_1 and ends when the energy falls below a second threshold θ_2. The advantage of an asymmetric threshold is that with a high start threshold θ_1 audio proto objects are not formed during pure noise periods. Since almost all sound signals have a sharply rising flank, but a slowly decaying signal energy toward the segment's end, the asymmetric thresholds allow capturing the rising flank and the tail of the segment, while excluding pure noise elements. The parameter θ_1 depends on the hardware characteristics and needs to be adapted to the background noise level. We set $\theta_2 = 0.6 \cdot \theta_1$. Although there are more refined methods for segmentation, our simple approach turned out to be sufficient for our purposes. Fig. 4 gives an example of the segmentation process.

This approach is currently limited to situations where speakers alternate without any overlap. Separating concurrent sounds in real-world applications has been demonstrated before (see e.g. [1], [2]). Our approach is targeted at scenarios where one or more interactors communicate with the robot in a cooperative way (similar to the concept of Motherese). The main challenge in such a scenario is that all types of environmental sounds may distract the robot from its main interaction partner, while concurrently active sources are not (yet) a major concern.

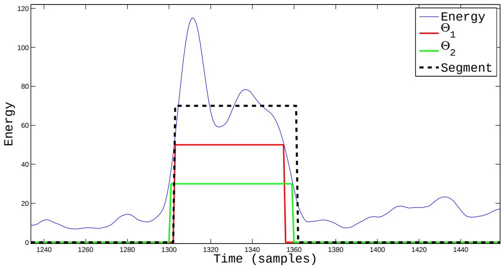

Fig. 4. Example of energy-based segmentation (dashed line) using two thresholds (upper threshold in red, lower in green). The energy was computed after stationary background noise was removed by spectral subtraction.

B. Feature Compression

The feature compression stage integrates 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. Energy is represented as the mean value over all samples in the segment (of length L).

$$P_{energy} = \frac{1}{L} \sum_{s \in S} A(s) \quad . \tag{3}$$

We also compute the difference of signal energy over time (increase and decrease of energy) and store this information in the response of a set of nodes representing different values. In total 11 nodes are used which respond to slope values of δA^r = [-5 -2 -1 -0.5 -0.1 0 0.1 0.5 1 2 5]. For every sample the closest matching node increases its response by one. This type of representation offers a constant size independent of segment length while still retaining some information about the distribution of δA values. The current setting of node centers is hand-crafted but could be learned from the statistics of proto objects.

The representation of the source location is the accumulated position evidence for all samples:

$$P_{position}(\alpha) = \sum_{s \in S} E(\alpha, s), \tag{4}$$

where α is the azimuth angle of the source, and $E(\alpha, s)$ the evidence for azimuth angle α in sample s. $E(\alpha, s)$ was integrated over time with a constant τ = 100 ms, see [3]. The same approach is employed for the spectral vector: $P_{spectral}(c) = \sum_{s \in S} A(c, s)$. An example audio proto object can be seen in Fig. 1.

IV. Results

We were using sounds 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 background sounds and directed speech and feature values for the two sound categories compared.

For each category the sound databases were divided equally into two sets, one for computing mean feature vectors and codebook vectors, and one for evaluation. The databases contained in total 317 environmental sounds and 82 speech commands recorded in a realistic scenario. Environmental sounds were measured by recording sound for approximately 9 min when the robot was turned on and people were working normally in the robot lab - including working on the robot's hardware and talking to each other. Robot-directed calls were recorded from four people at different positions calling the robot's attention repeatedly. Both recordings represent realistic, typical cases of audio signals of the two different types in our standard setting.

A. *Sound localization*

Our main application scenario of the audio proto objects approach so far is selective sound localization. We therefore investigated the performance of the proto object system in an offline localization scenario. We used a database of 29 sounds for 19 different relative horizontal positions between 90° and -90° recorded with ASIMO's microphones. The data was measured by producing sound from a stationary loudspeaker and turning the robot's head. A serious problem is that when the head turns to the sides, one of the microphones is very close to the robot's fans making sound processing extremely challenging. Using 14 sounds for training the remaining 15 were used for evaluating the precision of the localization.

Localization in the proto object framework differs from the conventional approach (as described in [3]) since the number and characteristics of the proto objects and therefore also the localization precision depends on the segmentation. Inevitable, some very short proto objects are generated which are difficult to localize correctly. In the standard approach other, more salient, parts of the sound dominate the localization response so that the problem could be neglected. To have a fair comparison we evaluated the proto object based localization for different settings of an energy-based filtering, i.e. responding only to proto objects with a certain minimal energy. As can be seen in Table I with increasing selectivity the localization precision increases substantially. Taking only the strongest 80% proto objects (*Top80*, this corresponds to approximately one proto object per sound file) the mean localization error falls to 4.5°. Considering that we are using only two microphones and operating in a large lab environment with substantial echo and background noise including the robot's fan noise, a localization error as low as 4.5° is very good.

For comparison our standard approach [3] has a mean azimuth localization error of 6.2°. Due to the high degree of fan noise at lateral head positions, we occasionally observed very large localization errors (up to 160°). The proto object based approach produced fewer localization outliers and with a setting of *Top50* or stricter no localization error larger than 30° is produced.

Filter	mean azi. error	perc. correct	> 30°
ALL	5.4°	61.7%	1.2%
Top99	5.1°	62.3%	0.6%
Top90	4.7°	63.1%	0.3%
Top80	4.5°	65.1%	0.4%
Top50	3.3°	71.3%	0.0%
Top20	2.5°	76.1%	0.0%
Standard	6.2°	72.6%	4.2%

TABLE I
LOCALIZATION PERFORMANCE WITH DIFFERENT SETTINGS OF ENERGY FILTERING (*TopXX* USES ONLY PROTO OBJECTS WITH A SIGNAL ENERGY THAT IS WITHIN THE TOP *XX* PERCENT OF ALL PROTO OBJECTS). THE RESULT IS GIVEN AS MEAN AZIMUTH LOCALIZATION ERROR, THE PERCENTAGE OF CORRECT LOCALIZATION (WITHIN 10° RECORDING PRECISION) AND TRIALS WITH MORE THAN 30° LOCALIZATION ERROR.

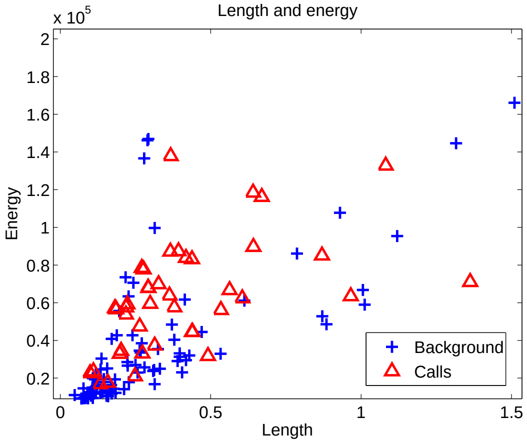

Fig. 5. Length (in seconds) and energy of audio proto objects for two types of sounds: background sounds and robot directed calls.

B. *Filtering based on segment length and energy*

After confirming the good localization performance we now turn to the problem of differentiating between background and call sounds. A first approach for filtering out background events is to analyze the mean segment length and signal energy of proto objects.

It turns out that most background sounds are rather short (mean 0.19 s compared to 0.36 s for calls) and have a low signal energy (mean 26000 compared to 62000 for robot directed speech). In Fig. 5 proto object length and energy are plotted. We used LVQ (*learning vector quantization* [14]) to learn optimal prototypes (background: length = 0.01 s, energy = 5300; calls: length = 0.4 s, energy = 71000) and then assigned each proto object to the better matching category. The result was that 94% of the background sound proto objects and 78% of the calls were categorized correctly.

Another obvious cue to separate different sound categories is sound source location. As we have seen, our approach allows a quite precise localization of sound sources. For some types of sounds the position might actually give a hint to the type of sound source. However, with our current approach, that is limited to the horizontal plane (azimuth),

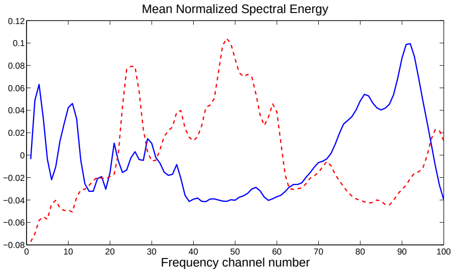

Fig. 6. Normalized mean spectral vector for background signals (solid, blue line) and speech (dashed, red line).

the position-related category information is not very high. It is the explicit purpose of the localization system to determine the position of the speaker wherever he or she is and turn to the corresponding position. We therefore should not make any assumptions on the position of the speaker. Similarly, background signals could potentially originate from all positions around the robot. If the sound localization is extended to elevation direction (as described in [12]) and distance, source position would actually be a more relevant cue (e.g. sounds from below are likely to be footsteps and very unlikely to be a human utterance).

C. Spectral energy similarity-based filtering

Since not all proto object features are scalar values, we extend our system by a similarity computation, that compares audio features between two proto objects and returns a scalar similarity value that can used for selecting the proper behavior. Fig. 6 visualizes the average spectral energies of proto objects, one for background signals and one for robot directed calls. Calls concentrate more energy in the middle frequency range while background sounds have more energy in the higher frequencies (a side effect of the GFB). We then compute the similarity of spectral vectors from proto objects of the evaluation set to these two reference vectors. Similarity is computed as the scalar product of normalized spectral energy vectors (zero mean, norm one). The result is shown in Fig. 7. Comparing the proto object's spectral vector to a reference vector for the two relevant categories, we could filter 86% of the background sounds, while keeping 93% of the relevant call proto objects.

D. Energy slope as a separating feature

We also investigated the potential of energy slope values for categorization. Similar to the approach for the spectral vectors we computed mean energy slope representatives for both sound categories (see Fig. 8) and calculated the similarity for all proto objects in both databases with the two reference vectors. Similarity was again based on the normalized scalar product. The result is shown in Fig. 9. Based purely on the energy slope distribution, background signals are to 74% correctly estimated as background and robot directed-speech with 95% as calls. The high success-rate could be partially due to the influence of signal energy on the slope values, but we also observed that speech signals tend to have a larger proportion of rising energy slope values than most environmental sounds.

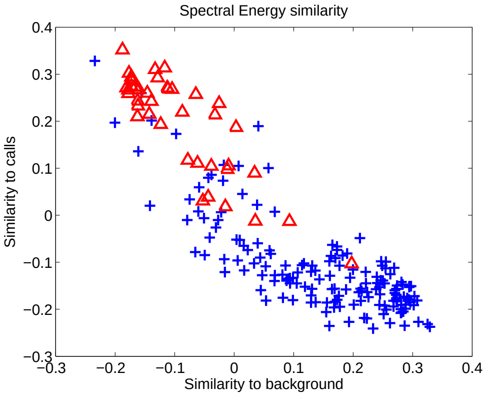

Fig. 7. Similarity of background sounds (*blue*, *'+'*) and robot directed calls (*red*, *'triangle'*) to the mean spectral vectors for background signals and calls.

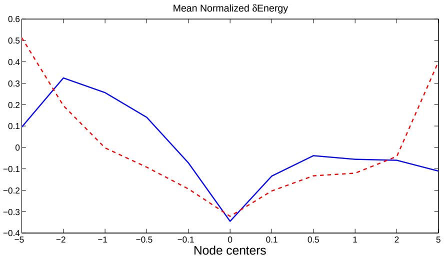

Fig. 8. Mean δA node response vector for background sounds (solid, blue line) and calls (dashed, red line).

E. Combination of filters

The combination of all cues into a single decision for selecting which category the current proto object belongs to and therefore the decision whether to attend to the sound or to ignore it provides an additional improvement compared to the single cues. Table. II summarizes the results for the individual cues and the combination of filters. Optimal prototypes vectors were again learned via LVQ using the training databases, while evaluation was performed on two separate sets of proto objects. Instead of δA node responses and spectral energy vectors we used the corresponding similarities to the representative vectors from the training database. In effect, we compute a 6-dimensional feature vector (length, energy, δA similarity to calls and background, spectral similarity to calls and background). Using LVQ we compute prototypical 6-dimensional feature vectors for both classes and then assign each proto object to the nearer class. In total 98% of the call and 91% of the background proto objects were categorized correctly. Since there is effectively a trade-off between the correctness for background and for calls, different vector quantization algorithms or parameters would probably result in different results along this pareto-front. We also note that the most basic approach, a mean energy-based filtering - comparable to an energy threshold

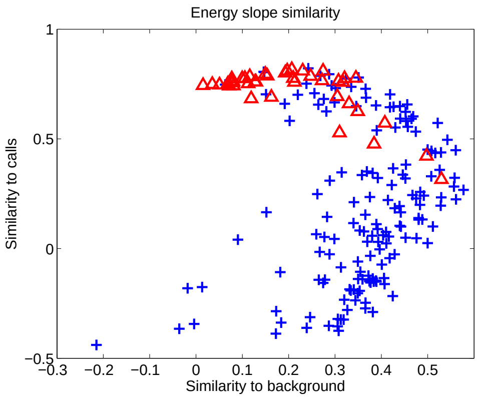

Fig. 9. Similarity of background sounds (*blue*,'+') and robot directed calls (*red*, *'triangle'*) to the mean energy slope vectors for background signals and calls.

Audio Feature	background correct	calls correct
Length	70%	80%
Energy	82%	76%
δA	74%	95%
Spectral	86%	93%
Length+Energy	94%	78%
All	91%	98%

TABLE II

PERCENTAGE OF CORRECT CATEGORIZATION OF PROTO OBJECTS AS BACKGROUND OR CALLS WHEN USING DIFFERENT AUDIO FEATURES.

operation, is on average 15% points worse than the combination of all filters.

F. Online system

We have successfully implemented the environmental sound filtering mechanism on our ASIMO robot as part of a larger integrated system similar to [4]. We are currently using only the filtering based on segment length and mean signal energy, but plan to extend the system to include all features as described above. 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 features, instead relying on relatively simple but robust audio cues. The selection of the proper behavior, i.e. to turn to the measured position of the current proto object or to ignore it, is made on the level of proto objects where filtering characteristics are easily adaptable. The complete sound processing system runs on a single standard multi-core machine in real-time.

V. SUMMARY AND OUTLOOK

The contribution we have made in this work is twofold: Firstly we have proposed a new concept for linking sound processing and behavior selection based on audio proto objects. These are formed by a segmentation process and combine a number of audio features into a compact representation that can be easily handled in higher processing stages. Secondly we demonstrated that very low-level features like mean signal energy and segment length can successfully distinguish between the two relevant sound categories in about 80% of the time. We could also show that using other features like spectral energy or energy slope, 91% of background signals can be selected out while 98% of the call proto objects are attended to. Due to the averaging over the full segment all computed features are quite robust to noise and echoes.

We plan to extend the concept of audio proto objects in several ways: segmentation also in the spectral dimension, addition of new audio features (like spectro-temporal features [15]), an adaptable analysis of proto objects, e.g. by changing filter thresholds or learning the characteristics of sound categories. We also need to unify the representation of audio features. It would be beneficial if all features could be represented in a similar way. A possible solution is to map scalar values to a population code representation like the one we used for the energy slope. It would then be possible to use a single approach for representing and comparing all features in the proto object. Along this line we also plan to improve the categorization, which is currently based on a rudimentary LVQ scheme.

In addition to filtering irrelevant sounds the concept can also be used to group similar proto objects or to integrate audio and visual proto objects for cross-modal learning or cue integration.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] N. Mesgarani, M. Slaney, and S. A. Shamma, “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 920–930, May 2006.

[6] N. S. Chu, S. and C.-C. J. Kuo, “Environmental sound recognition using mp-based features,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP*, 2008.

[7] M. Cowling and R. Sitte, “Comparison of techniques for environmental sound recognition,” *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003.

[8] W. W. Gaver, “How do we hear in the world?: Explorations in ecological acoustics,” *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

[9] ——, “What in the world do we hear?: An ecological approach to auditory perception,” *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

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

[11] 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.

[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.

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

[14] A. Sato and K. Yamada, “Generalized learning vector quantization,” *Advances in Neural Information Processing Systems*, vol. 7, pp. 423–429, 1995.

[15] 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.