Teaching a Humanoid Robot: Headset-Free Speech Interaction for Audio-Visual Association Learning

Martin Heckmann, Holger Brandl, Jens Schmüdderich, Xavier Domont, Bram Bolder, Inna Mikhailova, Herbert Janßen, Michael Gienger, Achim Bendig, Tobias Rodemann, Mark Dunn, Frank Joublin, Christian Goerick

2009

Teaching a Humanoid Robot: Headset-Free Speech Interaction for Audio-Visual Association Learning

Martin Heckmann[1], Holger Brandl[1,2], Jens Schmuedderich[1], Xavier Domont[1,3], Bram Bolder[1], Inna Mikhailova[1], Herbert Janssen[1], Michael Gienger[1], Achim Bendig[1], Tobias Rodemann[1], Mark Dunn[1], Frank Joublin[1], Christian Goerick[1]

Abstract— **Based on inspirations from infant development we present a system which learns associations between acoustic labels and visual representations in interaction with its tutor. The system is integrated with a humanoid robot. Except for a few trigger phrases to start learning all acoustical representations are learned online and in interaction. Similar, for the visual domain the clusters are not predefined and fully learned online. In contrast to other interactive systems the interaction with the acoustic environment is solely based on the two microphones mounted on the robots head. In this paper we give an overview on all key elements of the system and focus on the challenges arising from the headset-free learning of speech labels. In particular we present a mechanism for auditory attention integrating bottom-up and top-down information for the segmentation of the acoustic stream. The performance of the system is evaluated based on offline tests of individual parts of the system and an analysis of the online behavior.**

I. INTRODUCTION

While pursuing the goal of developing an autonomous intelligent system taking inspirations from infant development is a very promising road. Even though the balance between nature and nurture is an open issue, the strong role the interaction with its caregiver, i.e. nurture, plays in a child's development is undeniable. Translated to the development of intelligent systems it is clear that many of the abilities one expects from such a system should be learned in interaction.

Previously we developed a system which enabled our humanoid robot ASIMO to learn associations of predefined relative position clusters (”left”, ”right”, ...) with speech labels as well as associations between robot motions (”forward” and ”return”) with speech labels [1]. The two major improvements we present here are that the visual clusters are not predefined anymore but learned online and that the acoustic interaction is now completely based on the microphones mounted on ASIMO . The latter required the introduction of a model for auditory attention, i.e. which sounds to interpret and which not. Many approaches to improve the speech signal on robotic systems and models of auditory attention exist, but to our knowledge none of these was successfully integrated in a truly interactive system where the robot was moving and listening at the same time [2], [3], [4], [5]. Due to the unfavorable acoustic conditions on a mobile robot almost all current robotic systems use a headset mounted close to the speakers mouth when interacting with a robot [6], [7], [8] (see [2] for a headset free interaction in a stop-perceive-act paradigm).

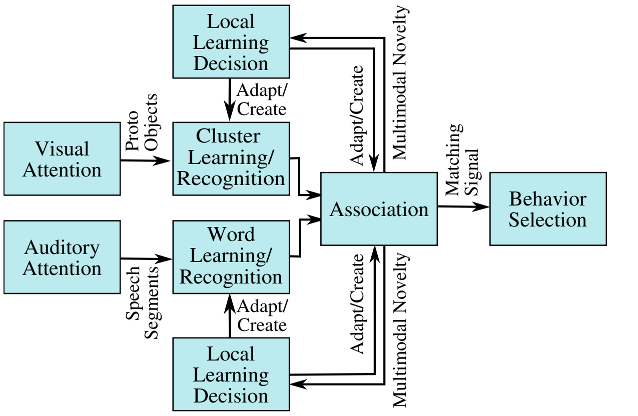


Fig. 1. Overview of the system with a focus on the sensory and representation parts. See [1] for details on the general architecture.

Fig. 1 shows an overview of our system where the focus is set on the sensory and representation parts. The general architecture is described in more detail in [1]. The following sections will describe the main building blocks, i.e. auditory and visual attention, online learning of speech labels and visual clusters, association building, and behavior generation. In the last sections we will evaluate sub-parts of our system on offline data and interpret the online behavior. To make the following details on the different modules more accessible we firstly describe the kind of human-robot interaction our system allows for.

II. INTERACTING WITH ASIMO

The design of our system targets on bootstrapping multimodal representations with minimal initial knowledge and enabling a continuous development by learning in interaction with a tutor. For instance our system can learn a cluster in the relative visual position space, an arbitrary speech label, and the association between both. To focus our system's attention to particular characteristics of a scene we use some phrases which can trigger a learning session, e.g. ”Learn

[1] Authors are with the Honda Research Institute GmbH, Offenbach/Main, Germany, firstname.lastname@honda-ri.de

[2] Holger Brandl is with the Research Institute for Cognition and Robotics, University of Bielefeld, hbrandl@cor-lab.uni-bielefeld.de

[3] Xavier Domont is with the University of Darmstadt, Institut für Automatisierungstechnik, FG Regelungstheorie & Robotik, xavier.domont@rtr.tu-darmstadt.de

where this object is.". A typical learning session consists of the following steps:

1) The tutor enters the interaction range of ASIMO so that it either sees the tutor or an object he is presenting.
2) The tutor utters one of the predefined learning phrases to teach categories as relative position, size, or a label to a movement of ASIMO.
3) The tutor presents an instance of the cluster to be learned, e.g. by showing and moving an object in the left field of view of ASIMO, while uttering the label he wants to associate to this cluster a few times (5-8).
4) When the tutor keeps silent for a few seconds the system ends the learning session and shows only reactive behavior.

To evaluate what the system has learned the tutor presents an object in one of the learned clusters and utters the associated label. If the active cluster and the recognized cluster do match ASIMO nods with its head. Otherwise ASIMO shakes its head and continues trying to find matches. If in a given time the match is found ASIMO finally nods and disables the expectation.

III. VISUAL ATTENTION

To cope with the concurring stimuli impinging continuously on our eyes and ears our brain disposes of mechanisms to selectively focus on only a few stimuli at the time, a process usually referred to as attention. Common models of attention, auditory or visual, comprise a stimulus driven bottom-up saliency stage and a top-down modulation to enhance or suppress certain types of stimuli [9], [10]. Our visual attention system is mainly bottom-up driven and based on the concept of proto-objects. Proto-objects are regions in the visual field that are formed by a common grouping feature, can be tracked over multiple images, and are stabilized both in space and time. Grouping features used are depth, proper motion, planar surfaces, and similarity to a given color (see [11] for more details). The visual scene description consists of a (possibly empty) set of possibly interesting entities that are close to the robot, move, are large planes, have a certain color, or any possible combination of these. From the set of proto-objects one is selected for interaction, i.e. ASIMO can point, walk, and gaze towards them. There is no need for any object recognition in order for ASIMO to start interacting, although it can be used to modify behaviors [12].

Mainly objects in the peri-personal range, i.e. very close to the robot and covering a large amount of its field of view, are represented as proto-objects. With these proto-object in its peri-personal range the robot does interact. The concept of peri-personal range reflects observations from the way small children perceive the world [13]. Additionally, the proto-object concept also covers visual stimuli in an inter-personal distance (here 1 - 2 m away). Their instantiation is solely based upon proximity, i.e. depth. They are not interacted with by the robot, but are used as top-down information for the auditory attention.

IV. AUDITORY ATTENTION

The purpose of our auditory attention mechanism is to decide based on bottom-up signal information and top-down modulation to which auditory events ASIMO should listen, i.e. segment them and transfer them to the recognition, and which to ignore (compare Fig. 2).

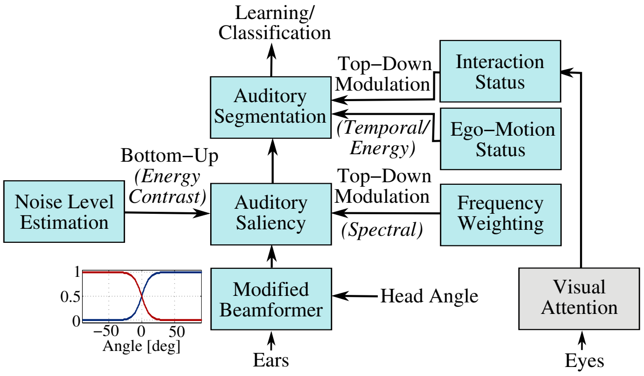


Fig. 2. Overview on the auditory attention model

When interacting without a headset with ASIMO a multitude of noise sources have to be dealt with. Stationary background noise from the computers in the room or the air-conditioning can be treated with conventional spectral subtraction methods. When ASIMO turns its head during interaction the microphones on ASIMO's head change their relative position to the noise generating fans in the backpack. This renders the noise of ASIMO's fans instationary. The noise emitted by the motors driving the arms and legs motion are not only instationary but due to their proximity to the ears of ASIMO they easily attain signal powers above those of the speech signal. Additionally there is the possibility of people speaking in the background. This can not be distinguished from the interactors voice based on spectral properties.

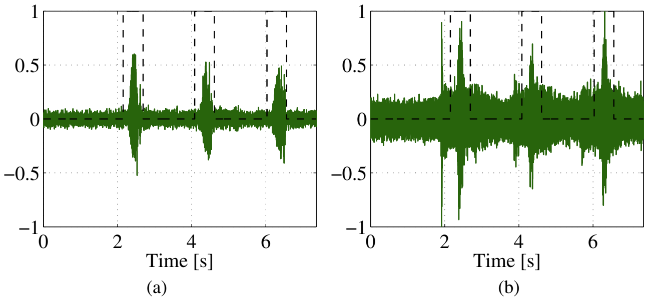


Fig. 3. Comparison of the same sound signal recorded during interaction with ASIMO once with a headset (a) and once with ASIMO's ears (b). The interactor is saying 3 times "left". The dashed lines indicate the detected speech segments.

When comparing the headset recording in Fig. 3a with the one from ASIMO's ears (Fig. 3b) the high noise floor due to the background noise and the fan noise can easily be distinguished. The high energy signal in Fig. 3b just before the first utterance of "left" of the interactor is the first movement of ASIMO's arm when it points to the interactor.

As can be seen from Fig. 3a in the headset recording this signal is barely audible.

A. Bottom-Up Saliency

The first step in the auditory bottom-up saliency is a contrast enhancement between the environmental noise and the speech signal based on a modified two channel delay and sum beamformer followed by an adaptive noise level estimation.

1) Modified Delay and Sum Beamformer: In normal interaction ASIMO looks to the object presented by the interactor, who typically presents the object in front of him. Hence one can assume that the speech signal is always coming from the looking direction of ASIMO. Therefore, when the head pan angle is small we use a delay and sum beamformer assuming a sound source at 0°, i.e. we add the signals form the left and right microphone. When ASIMO turns its head the *Signal to Noise Ratio (SNR)* in the two ears is dramatically different. For head pan angles of more than 20° it is better to use only the microphone farthest away from the fans. The transition between the two approaches is obtained via a continuous blending between the two ears depending on the head angle (compare Fig. 2).

2) Adaptive Noise Level Estimation: The basis of the noise estimation is an adaptation of the *Improved Minimum Controlled Recursive Averaging (IMCRA)* algorithm [14]. In contrast to the original implementation we transform the signal via a Gammatone filterbank into the frequency domain. The Gammatone filterbank constitutes a set of band-pass filters modeling the properties of the human cochlea. In the IMCRA algorithm the energy of the stationary parts of the acoustic signal, i.e. the background noise, are estimated and combined with the current signal energy to calculate an instantaneous speech probability for each filter-bank channel. The results of these contrast enhancement steps are depicted in Fig. 4 and constitute the bottom-up saliency signal.

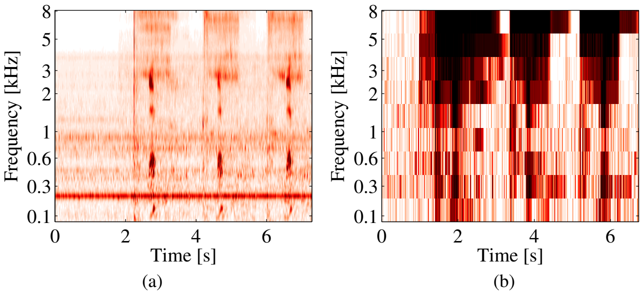


Fig. 4. Visualization of the contrast enhancement for the signal shown in 3.b. In (a) the signal is shown after application of the adaptive beamformer and transformation into the frequency domain via the Gammatone filterbank. The result of the contrast enhancement, a frequency dependent speech probability, is shown in (b). Dark colors indicate high probability.

B. Top-Down Modulation

After the bottom-up saliency calculation not only speech but all non-stationary signal parts are salient, including sounds produced by the movements of ASIMO (compare Fig. 3a and b and Fig. 4b). To suppress these noise sources additional top-down information is necessary to modulate the bottom-up saliency.

1) Spectral Modulation: The first form of top-down information we use is the spectral characteristics of the noise produced by ASIMO's movements. Arm and leg movement noise typically covers the speech signal for frequencies above 3.5 kHz. Additionally, leg movement noise has more energy than the speech signal for frequencies below 400 Hz. For the time being we only want to tune the auditory attention to speech signals. Therefore, we have chosen a frequency weighting of the bottom-up saliency which attenuates signals below 400 Hz and above 3.5 kHz. To obtain the modulated saliency signal the bottom-up saliency signal is multiplied with the frequency weighting and summed over all frequency channels. A threshold on this signal determines signal parts to be salient and hence a possible start of a speech segment.

2) Ego-Motion Status: Additionally we use the movement status of the robot to modulate the attention. We calculate the speed of the arm and leg motion and adapt the responsiveness, i.e. the speech segment detection threshold, of the attention system accordingly. The current setting allows the interaction via speech while ASIMO is moving its arms or makes small steps. However, when it walks or in the brief but very noisy instant when it starts raising the arm from the rest position it will only detect speech when shouted at.

3) Interaction Status: Another very important top-down information we recruit is the current interaction status of ASIMO which we determine based on the visual attention system. When ASIMO neither sees an object in its peri-personal space or a human in its inter-personal space it assumes that nobody is interacting with it and hence it raises the minimal activity threshold for its auditory attention. Currently the threshold is raised up to a level where it is not able to detect speech segments anymore. With this mechanism the voices of people standing in the background can be suppressed in non-interaction phases.

4) Minimum Segment Length: The final top-down modulation factor we use is the minimum segment length. Many sound events we want to ignore when focusing on speech are rather short, e.g slamming of a door or something dropping to the floor, and therefore can be rejected based on a minimum length criterion. We use a minimum segment length of 110 ms. When we detect an activity in the modulated saliency we accumulate the evidence for this time span and decide at the end if we accept this as the start of a speech segment which will be continued or if we reject it. Due to the latency introduced in the overall system it is not advisable to prolong the accumulation time much more even though this would allow to reject more erroneous segments. As a result of the rather long reverberation time in our robotics laboratory ($\tau_{60} = 810\,\mathrm{ms}$) this minimum segment length contributes only to a smaller extend to the overall system performance.

The segmentation of the speech signal resulting from the combination of the bottom-up saliency and the speech oriented top-down modulation is visualized in Fig. 3b. As can be seen the signal parts resulting from the arm movements do not trigger the start of the segment.

V. Acoustic Feature Extraction

The acoustic feature extraction is continuously running and the segmentation obtained by the auditory saliency only gates these features. As features we use a combination of RASTA-PLP features [15] and the HIST features developed by ourselves. We could show previously that this combination yields in the order of $20 - 40\%$ better recognition performance when compared to RASTA-PLP alone [16]. However, it is very difficult to measure this improvement for an online system in free interaction.

VI. ONLINE LEARNING

Initially the system has only very little knowledge. The visual clusters and the speech labels are fully learned in interaction. To ease the use of our system, we have favored a focused attention mechanism over unconstrained associative learning. Learning of new clusters is only possible during so called *learning session*s, that allow the different perceptual modalities to accumulate sufficient samples for cluster learning. Currently such sessions are triggered by uttering a predefined criterion that constraints the non-speech attention of our system to a certain visual object property (like its size) or robot action. Within a session an object with the property to be labeled is presented, and matching speech labels are uttered several times. After a session has timed out, speech and the visual subsystem in focus determine the novelty of the current session to existing clusters. For each pair of two associated clusters a weighted summation of their activations is performed, forming a multimodal novelty signal. These signals are returned to their originating classifiers which individually decide whether the session data should be represented by a new cluster or whether the best matching cluster should be adapted. We call this a *local learning decision*. Finally, newly created clusters are associated with each other. The visual cluster learning, association learning, and behavior organization will only be explained to a level so that the reader can grasp the overall behavior of the system. The focus is on the auditory aspects. The remaining parts will be detailed elsewhere.

A. Online Word Learning

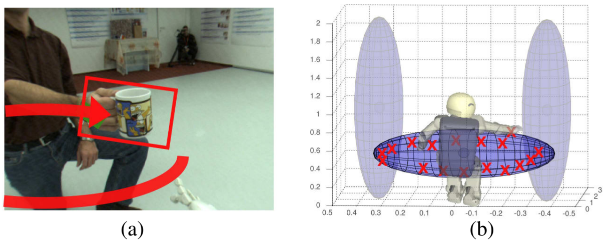
We apply Hidden Markov Models for speech representation, and the features described in Sec. V. Each speech cluster is modeled as an 8 state HMM with Bakis-topology. According to the local learning decision, either a new speech model is learned or the best matching speech cluster is updated. New speech clusters are initialized with the best matching label model, and are subsequently estimated using segmental k-means training with the collected session samples. If the target class in the teaching signal is already modeled, the according speech cluster is updated with maximum a-posteriori training.

During decoding we use a combined search space that includes HMM-subgraphs of already acquired label models, the above-mentioned predefined learning-criteria, and a generic background model learned prior in interaction as described in [17]. The latter equips our system with the ability to reject *Out Of Vocabulary (OOV)* utterances. Decoding results are accordingly split into commands used to trigger the learning sessions and recognized labels. The latter are combined to an activation vector which is passed to the online association learning.

Speech novelty for a training session $\mathcal{S} = s_1, \cdots, s_N$ containing N samples of an auditory label is calculated as follows. First, we determine the model $\hat{\lambda}$ that best matches to the session data as the model that maximizes the session joint liklihood $P(\mathcal{S}|\hat{\lambda})$. Next, we approximate the probability density function $p_{\mathcal{H}}(l_{\hat{\lambda}}(s))$ with a Parzen-model with Gaussian kernels estimated on the segments contained in $\mathcal{H}$. Hereby $l_{\hat{\lambda}}(s)$ denotes the likelihood of a segment s given the model $\hat{\lambda}$ and $\mathcal{H}$ the set of segments used for the estimation of $l_{\hat{\lambda}}$. In the same fashion we estimate the probability density function $p_{\mathcal{S}}(l_{\hat{\lambda}}(s))$ based on the likelihoods of each element in the training session $\mathcal{S}$. Finally, a session is considered to contain samples of a new, not yet represented word if $q_{0.3}(p_{\mathcal{S}}) > q_{0.5}(p_{\mathcal{H}})$. Hereby $q_{\alpha}(p)$ denotes the α-quantile of a probability density function p.

B. Online Visual Cluster Learning

For learning of visual properties different features of the currently focused proto-object are used, such as a vector of its 3d position in heel coordinates or the absolute value of its 3d size in camera-coordinates. However, the underlying classifiers are identical.
Each cluster is represented by a multi-dimensional Gaussian, consisting of a cluster-center and a covariance. The activation of each cluster given some feature-vector is based on the distance between the cluster-center and the feature vector, integrated over time. The larger the distance, the lower the cluster activation (refer to [11] for more details).



(a) (b)

Fig. 5. Image taken from ASIMO's camera during learning of "bottom" in (a) and illustration of the construction of the clusters in (b).

At the end of each learning session, the feature-vectors accumulated during the session are used to iteratively update the mean value and the covariance of the cluster specified by the local learning decision (compare Fig. 5). If the learning decision indicates the creation of a new cluster, that cluster is initialized with the mean value and covariance computed from the collected data.

C. Online Assocition Learning

Initially, the system neither contains any clusters nor associations. The learning of new associations assumes synchronously presented clusters in two different modalities to belong together. Therefore, the local learning decisions of the speech and the visual classifier in focus can be used to define the mapping between the two modalities. If both classifiers vote for the creation of a new cluster these two clusters are associated with each other. In the case where only one learning decision demands the creation of a new cluster this new cluster is associated with the already existing one in the other modality.

D. Behavior Organization

The system's behavior is organized in several parallel layers of control (see [1] for more details). The lowest level serves the whole body motion control of the robot, including a basic conflict resolution for different target commands and a self collision avoidance of the robot's body. A reactive behavior control implements task-unspecific interaction with the environment. It is based on tracking of proto-objects as sensory input and uses a competitive dynamics for arbitration of about 20 alternative behaviors like point, grasp, gesture and approach. Multiple behaviors can be active at any time e.g. pointing with one hand while gesturing with another, but behaviors also compete for execution if they are mutually exclusive, e.g. nodding and shaking the head. The behavior selection is based on two values for each behavior, a fitness that describes the applicability of the behavior given the environment and robot state and an external bias that allows to suppress or boost specific behaviors by other control layers. On top we built a layer that allows for an expectation-driven behavior. In this implementation the speech recognition generates the expectation of the visual properties or the activated behaviors. In case of expectation match the robot nods using the bias mechanism. Otherwise, for a fixed period of time the robot tries to resolve the mismatch by sending an appropriate top-down signal to the reactive layer.

VII. Results

First, we will present results we obtained from offline tests. After this we will discuss the online performance of our system.

A. Offline Tests

The first issue we want to investigate is how the word recognition performance depends on the number of training samples presented in the learning session. Therefore, we recorded a database where our interactor uttered 21 different words (e.g. "left", "right", "top", ...) each 20 times. While doing so he was standing in our robotics laboratory (reverberation time $\tau_{60} = 810\,\mathrm{ms}$) in front of the robot which was turned on but not moving. Hence the recoding conditions were very close, but due to the passive robot not identical, to the ones faced in the interaction. In Fig. 6a it can be seen that the recognition performance strongly depends on the number of training samples used. Each *Word Error Rate (WER)* value represents the mean of a 10-fold cross-validation. The bars indicate the minimum and maximum value in each validation step. From 7 examples on the performance is by far sufficient to allow for a smooth interaction. Reasons for the good recognition scores we see are certainly the quite small vocabulary ($\approx$ 20 words) and the fact that we train and test under the same conditions. It is well known that such matched training has a much larger effect than most preprocessing methods.

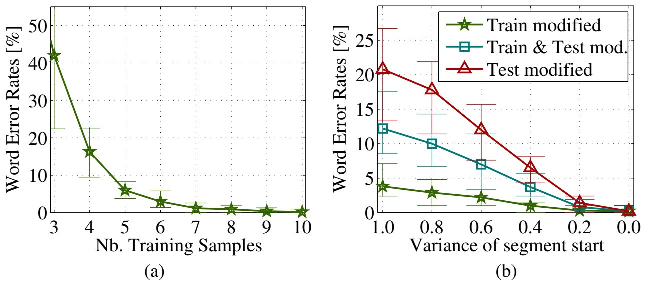


Fig. 6. Word Error Rate (WER) results when the training size was varied (a) and when the segment boundaries were changed (b).

The next question we want to address is the influence of the segmentation of the speech signal on the recognition performance. The segmentation based on the auditory attention can be erroneous either in the learning phase, the testing phase or both. We investigated this by randomly varying the detected segment start and stop boundaries. To avoid cutting off parts of the speech signal segments were only prolonged relative to the originally detected boundaries. Hence this test reflects the sensitivity of the learning and recognition algorithms on additional background noise at the start and end of the segment. We altered the segment start and stop points with noise from a folded Normal distribution, i.e. $Y \sim |N(0, \sigma^2)|$, with varying variances. As can be seen from Fig. 6b the word error rates increase rapidly with increasing variance. The tests are based on 10 training samples and again a 10-fold cross-validation. Alterations of the segments only in the training phase has only a rather small impact on performance. From this we conclude that the learning algorithm can cope quite well with additional noise at the beginning and end of the segment which can be due to the averaging over 10 segments in the learning phase. However, when the segment boundaries are also altered in the testing the performance decreases significantly. We obtain the worst results if the segment boundaries are only altered in the testing. In this case already a variance of 0.2 increases the error rates 7 times (from 0.2% to 1.4%, compared to 0.8% with modifications in training and testing and 0.3% only in training). The results clearly demonstrate the importance of the auditory attention system and the need for correct segmentation of the audio stream.

Finally we evaluated the performance of the novelty detection. As outlined in section VI-A its performance depends on the accuracy of the estimated likelihood-distributions. To assess the quality of our session-based novelty method, we evaluated changes in the size of $\mathcal{S}$ and $\mathcal{H}$. We computed F_1- measure for each configuration as the equally weighted harmonic mean of precision and recall. Hereby, a true positive was defined to be a successful detection of an known word as known, and a true negative was counted when an unknown segment was correctly identified as new. Results are shown in figure 7. As expected the performance of our

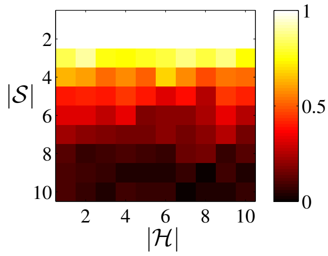


Fig. 7. Performance of the speech novelty algorithm evaluated for different reference-session and test-session sizes.

system improves with increasing trainings- and test-sessions. Surprisingly, the influence of the test-session was found to be much stronger.

B. Online Results

The methods described above were integrated in an online system running on Honda's humanoid robot ASIMO. Based on the mentioned pre-trained key-phrases the system is able to learn visual and speech clusters in an interaction pattern as described in Sec. II. In this interaction also the learning of synonyms for previously learned speech labels is possible. We tested this e.g. while learning the size of a book and a jug. The two objects have a similar size but a quite different appearance. Hence the visual novelty is low as the representation is based on the object size and not its identity. Due to the correspondingly high auditory novelty the visual cluster is adapted and a novel speech model is generated and associated with the original visual cluster. Therefore we could teach the English word "large" when presenting the book and then adapt the visual model and at the same time learn a speech synonym while presenting the jug and uttering "oki", the Japanese word for large. When we then evaluate what was learned by showing the book and uttering "oki" the system successfully associates the active visual cluster and the speech label.

VIII. CONCLUSION

We presented a system enabling to teach our humanoid robot ASIMO associations between visual clusters and speech labels in natural interaction. The system continues our previous work [1]. Novel aspects are that the speech interaction is solely based on the microphones mounted on ASIMO and that the visual clusters are learned from scratch. To our best knowledge this is the first interactive robotic system without headset not following a strict stop-perceive-act paradigm. This required the development of a system for auditory attention so as to reduce the number of misclassifications to a minimum. Our attention system integrates different bottom-up and top-down cues. None of these cues by themselves would be powerful enough but via integrating them we are able to obtain a robust segmentation of the speech signal allowing for online learning and recognition of the labels. The interaction without a headset significantly added to the naturalness of the interaction as in principle anybody can just step up to the robot and interact with it. This is only limited by the fact that the key phrases triggering the learning session are pre-trained in a speaker dependent fashion. Nevertheless, it is still rather easy to trick the system by producing other sounds which will erroneously be identified as speech. Therefore, further developments of the system will focus on the integration of spectral features in the segmentation process and improvements of the novelty detection/rejection mechanisms in the recognition stage. Additionally, a better detection of humans and their interaction state, e.g. their gaze direction, will also be necessary to further increase the naturalness of the interaction.

REFERENCES

[1] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmuedderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *IEEE-RAS Int. Conf. on Humanoid Robots*. 2008, IEEE-RAS.

[2] E.S. Neo, T. Sakaguchi, and K. Yokoi, "A natural language instruction system for humanoid robots integrating situated speech recognition, visual recognition and on-line whole-body motion generation," in *IEEE/ASME Int. Conf. on Advanced Intelligent Mechatronics (AIM)*, 2008, pp. 1176–1182.

[3] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H.G. Okuno, "Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation," in *Proc IEEE/RSJ Int. Conf. on Robots and Intell. Syst. (IROS)*, 2008, pp. 1718–1723.

[4] Y. Takahashi, H. Saruwatari, and K. Shikano, "Real-time implementation of blind spatial subtraction array for hands-free robot spoken dialogue system," in *IEEE/RSJ Int. Conf. on Intel. Robots and Systems (IROS)*, 2008, pp. 1687–1692.

[5] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," *IEEE Trans. on Neural Networks*, vol. 15, no. 5, pp. 1151–1163, 2004.

[6] J. Schmidt, N. Hofemann, A. Haasch, J. Fritsch, and G. Sagerer, "Interacting with a mobile robot: Evaluating gestural object references," in *Proc IEEE/RSJ Int. Conf. on Robots and Intell. Syst. (IROS)*, Nice, France, 22/09/2008 2008.

[7] R. Stiefelhagen, H. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling Multimodal Human–Robot Interaction for the Karlsruhe Humanoid Robot," *IEEE Trans. on Robotics*, vol. 23, no. 5, pp. 840–851, 2007.

[8] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Natural deictic communication with humanoid robots," in *Proc IEEE/RSJ Int. Conf. on Robots and Intell. Syst. (IROS)*, San Diego, 2007, pp. 1441–1448.

[9] J. B. Fritz, M. Elhilali, S. V. David, and S. A Shamma, "Auditory attention–focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437 – 455, 2007, Sensory systems.

[10] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–204, 2001.

[11] J. Schmuedderich, H. Brandl, B. Bolder, M. Heracles, H. Janssen, I. Mikhailova, and C. Goerick, "Organizing multimodal perception for autonomous learning and interactive systems," in *IEEE-RAS Int. Conf. on Humanoid Robots*. 2008, IEEE-RAS.

[12] C. Goerick, B. Bolder, H. Janssen, M. Gienger, H. Sugiura, M. Dunn, I. Mikhailova, T. Rodemann, H. Wersing, and S. Kirstein, "Towards incremental hierarchical behavior generation for humanoids," in *IEEE-RAS Int. Conf. on Humanoids*. 2007, IEEE.

[13] C. Yu, L.B. Smith, and A. Pereira, "Grounding Word Learning in Multimodal Sensorimotor Interaction," in *Proc. of the 30th Annual Meeting of Cognitive Science Society (CogSci)*, Washington DC, USA, 2008.

[14] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 5, pp. 466–475, 2003.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

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

[17] M. Vaz, H. Brandl, F. Joublin, and C. Goerick, "Learning from a tutor: Embodied speech acquisition and imitation learning," in *Proc. Int. Conf. Development and Learning (ICDL)*, 2009.