An Audio-Visual Attention System for Online Association Learning

Martin Heckmann\textsuperscript{1}, Holger Brandl\textsuperscript{1,2}, Xavier Domont\textsuperscript{1,3}, Bram Bolder\textsuperscript{1}, Frank Joublin\textsuperscript{1}, Christian Goerick\textsuperscript{1}

\textsuperscript{1}Honda Research Institute GmbH, Offenbach/Main, Germany
\textsuperscript{2}Research Institute for Cognition and Robotics, University of Bielefeld
\textsuperscript{3}Technische Universität Darmstadt, Control Theory and Robotics Lab

firstname.lastname@honda-ri.de, hbrandl@cor-lab.uni-bielefeld.de, xavier.domont@rtr.tu-darmstadt.de

Abstract

We present an audio-visual attention system for speech based interaction with a humanoid robot where a tutor can teach visual properties/locations (e.g. “left”) and corresponding, arbitrary speech labels. The acoustic signal is segmented via the attention system and speech labels are learned from a few repetitions of the label by the tutor. The attention system integrates bottom-up stimulus driven saliency calculation (delay-and-sum beamforming, adaptive noise level estimation) and top-down modulation (spectral properties, segment length, movement and interaction status of the robot). We evaluate the performance of different aspects of the system based on a small dataset.

Index Terms: attention, audio-visual, interaction, speech recognition, speech features

1. Introduction

Based on our previous work we developed a system which enables our humanoid robot ASIMO to learn associations of relative position clusters (“left”, “right”, ...) or object properties (“small”, ...) with arbitrary speech labels (see [1, 2] for more details). We use some predefined key phrases to trigger a learning session, e.g. “Learn where this object is.” Typically such a learning session starts with the tutor entering the interaction range of ASIMO and presenting an object. After triggering a learning session 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). When the tutor keeps silent for a few seconds the system terminates the learning session. To evaluate what the system has learned the tutor presents an object in one of the learned clusters and utters the associated label. With nodding or shaking the head ASIMO indicates if the visual and speech cluster do match.

The speech signal is solely captured by the microphones mounted on the robots head. This required to extend the existing visual attention system by a model for auditory attention. 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 [3, 4, 5]. Due to the unfavorable acoustic conditions on a mobile robot basically all current robotic systems use a close-talk microphone when interacting with a robot [6, 7, 8].

The following sections will describe the main building blocks of the combined audio-visual attention system and the online learning of speech labels. Finally we will present results of sub-parts of our system on offline data and interpret them.

2. Audio-Visual Attention

Attention allows us to selectively concentrate on one aspect of the environment while ignoring other things. Models of attention, auditory or visual, typically comprise a stimulus driven bottom-up saliency stage and a top-down modulation to enhance or suppress certain types of stimuli [9, 10].

In the following we will only describe those parts of our audio-visual attention system which are recruited to decide to which auditory events ASIMO should listen, i.e. segment them and transfer them to the recognition, and which to ignore (compare Fig. 1). Details on the role the visual part of this system plays in the organization of the behavior of ASIMO can be found in [1].

In an interactive scenario with a long distance between the speaker and the microphones on the robot a multitude of noise signals overlay with the speech signal. Hereby especially the noise generated by the robot plays an important role. It is instationary and due to its proximity to the microphones in the robot’s head easily attains signal levels above those of the speech signal (see Fig. 2). This includes the noise generated by its arm and leg movement but also the noise emanating from its cooling fans mounted on its back, as head movements change the relative position of the microphones to the fans.

2.1. Bottom-Up Saliency

In the bottom-up stage the contrast enhancement between the environmental noise and the speech signal is mainly achieved by reducing the background noise.

2.1.1. Modified Delay and Sum Beamformer

In a typical interaction ASIMO looks to the object presented by the interactor. Hence one can assume that the speech signal is

![Figure 1: Overview on the attention model](image-url)
2.2. Top-Down Modulation

In addition to speech also the instationary sounds produced by the movements of ASIMO are still salient after the bottom-up saliency calculation (compare Fig. 2 a and b and Fig. 3 b). To suppress these additional top-down information is necessary to modulate the bottom-up saliency.

2.2.1. 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 salient parts to be salient and hence a possible start of a speech segment.

2.2.2. Ego-Motion Status

We also use the movement status of the robot to modulate the attention. The responsiveness, i.e. the speech segment detection threshold, of the attention system is varied depending on the arm and leg speed. 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.

2.2.3. Interaction Status

Another very important top-down information we recruit is the current interaction status of ASIMO which we determine based on the visual part of the attention system. The visual part is mainly bottom-up driven and based on the concept of proto-objects, regions in the visual field that are formed by a common grouping feature as e.g. depth (see [12] for more details). One class are proto-objects in its peri-personal range, i.e. very close to the robot and covering a large amount of its field of view. With these proto-objects ASIMO does interact. A second class of proto-objects cover an inter-personal range (here 1 - 2 m away). Proto-objects in this range are assumed to be due to a human in interaction range. When no proto-object is present in the peri-personal or inter-personal space ASIMO assumes that nobody is interacting with it and hence 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 and hence in non-interaction phases voices of people standing in the background can be suppressed.

2.2.4. Minimal Segment Length

Most intruding sound events, e.g. slamming of a door, are rather short. Therefore, we use a minimum segment length (110 ms) as final top-down modulation factor. Activity in the modulated saliency is accumulated for this time span. Only when it surpasses the activity threshold a speech segment is started. The minimum length used is a trade off between the latency introduced hereby in the overall system and the potential to reject more erroneous segments. Due to the long reverberation time in our robotics laboratory ($\tau_{60} = 810$ ms) the 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. 2b. As can be seen the signal parts resulting from the arm movements do not trigger the start of the segment.
In focus determine the novelty of the current session to exist-

After a session has timed out, speech and the visual subsys-
tem to determine if a new matching label model, and subse-
sequently 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 [15]. The latter equips our system with the ability to reject unknown (Out Of Vocabulary (OOV)) utterances. Decoding results are accordingly split into commands used to trigger the learning sessions and recognized labels.

### 3. 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 [13] and the HIST features developed by ourselves (see [14] for details).

HIST features comprise two hierarchical levels: The first extracts local features and the second integrates them to more complex features, spanning the whole frequency range. The extraction of local features on the first level is performed via a 2D filtering with a set of 8 receptive fields. They have been learned using Independent Component Analysis on 3500 randomly selected local 16 × 16 patches on spectrograms preprocessed via a formant enhancement step using pre-emphasis and filtering along the frequency axis. On the second level we learn 50 filters with Non-Negative Sparse Coding on the responses of the filters of the first level. These filters span the whole frequency range and 40 ms in time. Delta (resp. double-delta) features were computed. Finally, the dimensionality was reduced from 150 to 39 using Principal Component Analysis.

To simulate the conditions of our interactive scenario signals where convolved with a room impulse response measured in our laboratory and noise recorded from ASIMO while not moving was added for the learning of the features. As dataset we used TiDigits. The results in Table 1 match with those presented in [14]: in their current development state HIST features perform less well than RASTA-PLP features, but improve the recognition performance when combined with the latter. This improvement is observed in the matched case (noise recorded on a resting robot) as well as when the noise added to the test set was recorded when the robot moves his arms or legs.

### 4. Online Learning

The purpose of the previously detailed auditory attention system is to enable online learning of visual clusters and corresponding speech labels. Visual clusters can e.g. be regions in the relative position space of the robot as “left” or “right” (see [12, 2] for details).

The utterance of a predefined key-phrase triggers the learning. 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. This information is used to determine if a new cluster/speech label has to be learned or if rather an existing representation should be updated.

For learning and recognizing the speech labels we apply Hidden Markov Models and the features described in Sec. 3. Each speech cluster is modeled as an 8 state HMM with Bakistopology. According to the 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 subsequently estimated using segmental k-means training with the collected session samples. If the target class in the teaching signal is already modeled, the stronger dependency on the number of training samples presented in the learning session. We recorded a small database where our interac-
tor was standing in our robotics laboratory (reverberation time $\tau_{10} = 810$ ms) in front of the turned on but not moving robot uttering 21 different words (e.g. "left", "right", "top", . . . ) each 20 times. Hence the recoding conditions were very close, but due to the passive robot not identical, to the ones faced in the interaction. As can be seen in Fig. 4 a from 6 training samples on the performance is by far sufficient to allow for a smooth interaction. The combination of RASTA-PLP and HIST shows a

### 5. Results

First we investigate the dependence of recognition performance on the number of training samples presented in the learning session. We recorded a small database where our interac-
tor was standing in our robotics laboratory (reverberation time $\tau_{10} = 810$ ms) in front of the turned on but not moving robot uttering 21 different words (e.g. "left", "right", "top", . . . ) each 20 times. Hence the recoding conditions were very close, but due to the passive robot not identical, to the ones faced in the interaction. As can be seen in Fig. 4 a from 6 training samples on the performance is by far sufficient to allow for a smooth interaction. The combination of RASTA-PLP and HIST shows a

<table>
<thead>
<tr>
<th>Noise type</th>
<th>fan noise</th>
<th>arm noise</th>
<th>leg noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASTA-PLP</td>
<td>32.4</td>
<td>35.2</td>
<td>40.2</td>
</tr>
<tr>
<td>HIST</td>
<td>56.3</td>
<td>71.8</td>
<td>70.0</td>
</tr>
<tr>
<td>RASTA-PLP + HIST</td>
<td>29.0</td>
<td>31.6</td>
<td>39.1</td>
</tr>
</tbody>
</table>

Table 1: Word error rates on TiDigits. Training was done with fan noise added and tests were performed in this condition or when noise from the robot’s arm or leg movements was added.

![Figure 4: Word error rates when the training size was varied (a) and when the segment boundaries were changed (b).](image-url)
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.

In the final test we investigated the impact of robot motion noise on the performance, an important aspect in our interactive scenario. We recorded another small dataset with our tutor uttering the labels (10 repetitions each) while the robot was turned off. To these recordings we added the noise generated by the robot while moving its arms or legs (while “stamping” on the spot). The recording of another database was necessary as recording the robots motions unavoidably also includes the fan noise. Hence adding the motion noise to the first dataset would result in twice the fan noise in the signal. We performed two tests. One where the segmentation was based on the energy signal prior to mixing with the noise but used the noisy signal for the recognition. This situation simulates a correct segmentation of the signal but is more susceptible to errors in the segmentation and hence, the attention system and the HIST features complement each other very well.

7. Acknowledgment

Many people of the Honda Research Institute contributed to this work. We want particularly to thank Jens Schmuedderich, Inna Mikhailova, Herbert Janssen, Tobias Rodemann, Michael Gieninger, Achim Bendig, and Mark Dunn.

8. References


