Acoustic Packaging

Lars Schillingmann[1] – Britta Wrede[2] – Katharina Rohlfing[3]

[1 2]Applied Informatics Group, Faculty of Technology, Bielefeld University, Germany

[1 2 3]Research Institute for Cognition and Robotics, Bielefeld University, Germany

[3]Emergentist Semantics Group, Center of Excellence, Cognitive Interaction Technology, Bielefeld University, Germany

Universität Bielefeld

CoR-Lab Research Institute for Cognition and Robotics

italk

Acoustic Packaging – Key Ideas

- Acoustic packaging makes use of the synchrony between the visual and audio modality in order to detect temporal structure in actions that are demonstrated to children and robots [1, 2].
- Support for action learning in robots
  - Acoustic packages form early units for further learning processes.
  - Feedback generation during tutoring.

Figure: A test subject showing how to stack cups to an infant [3].

Related Work

- Proposed and termed acoustic packaging by Hirsh-Pasek & Golinkoff [1]
  - Language helps to divide a sequence of events into units.
- Study by Brand and Tapscott [2]
  - Co-occurring infant-directed-speech and motion helps infants to group sequences of movement into meaningful units.

System Requirements

- Segmentation: A temporal segmentation for at least one acoustic and one visual cue is required.
- Temporal synchronization: Visual and acoustic segments need to be temporally aligned.
- Timestamp concept: Helps in aligning segments created by different processing modules.
- Extensible: It should be easy to integrate further cues and modules that perform further processing towards learning.
- Online capable: A socially interactive robot should give feedback during tutoring.
- Visualization: A tool for inspecting and debugging the involved cues and the acoustic packaging process is needed.

System Overview

- Modular and decoupled approach
  - Modules communicate through a central memory the Active Memory [4].
  - The Active Memory notifies components about event types they have subscribed to.
- The audio signal is processed using the ESMERALDA speech recognizer.
- The visual signal is processed with the help of a graphical plugin environment called icewing.

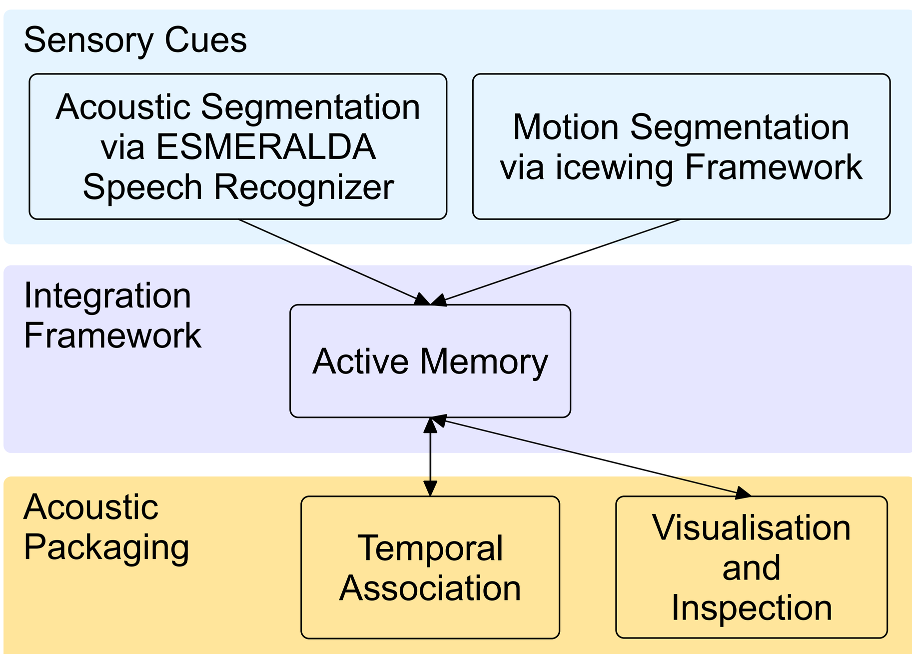

Figure: System overview with highlighted layers and their relation to the acoustic packaging system.

Acoustic Segmentation

- Segmentation into speech and speech pauses
- More robust in noisy environments than a simple voice activity detection.
- The speech recognizer is configured for monophoneme recognition.
- Phonotactics are modeled statistically via an *n*-gram model.
- Phoneme hypotheses and the audio signal are inserted in the Active Memory.
- The recognition process is incremental: Hypotheses are updated continuously.

Visual Action Segmentation

- Segmentation into motion peaks
- A peak ranges between two local minima in the amount of change in the visual signal.
- The amount of change is calculated by summing up a motion history image at each time step.
- Peaks are detected within a sliding window.

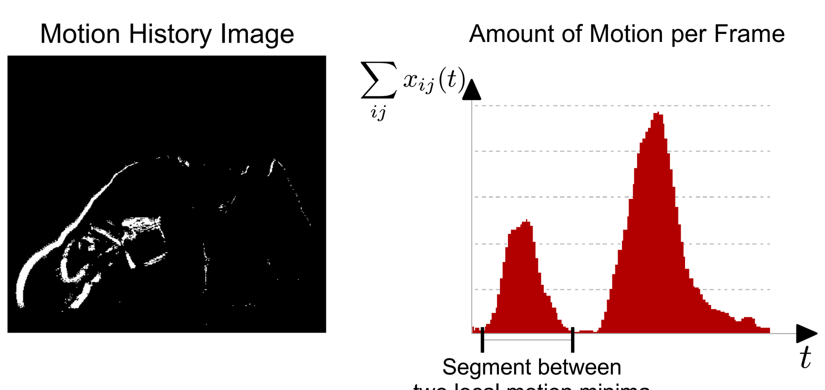

Figure: Motion History Image / Motion Peaks

Temporal Association

- This module maintains a timeline for acoustic and visual segments.
- Overlapping speech and visual segments are associated to one acoustic package.
- Acoustic packages are updated if the corresponding hypotheses from the signal processing modules are updated.

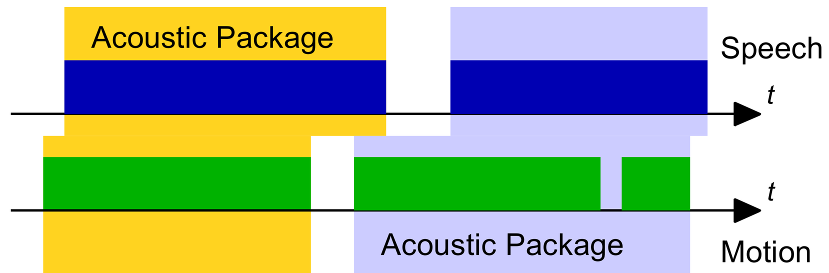

Figure: Illustration of the temporal association process. The example shows two acoustic packages.

Inspection and Cue Visualization Tool

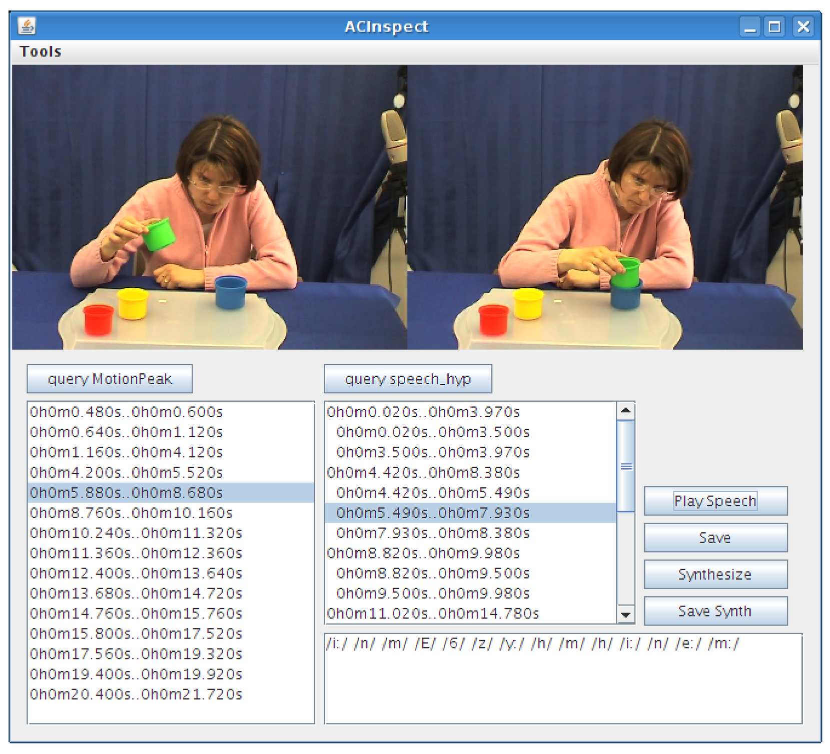

Figure: Inspection tool showing the beginning and end frame of a motion peak.

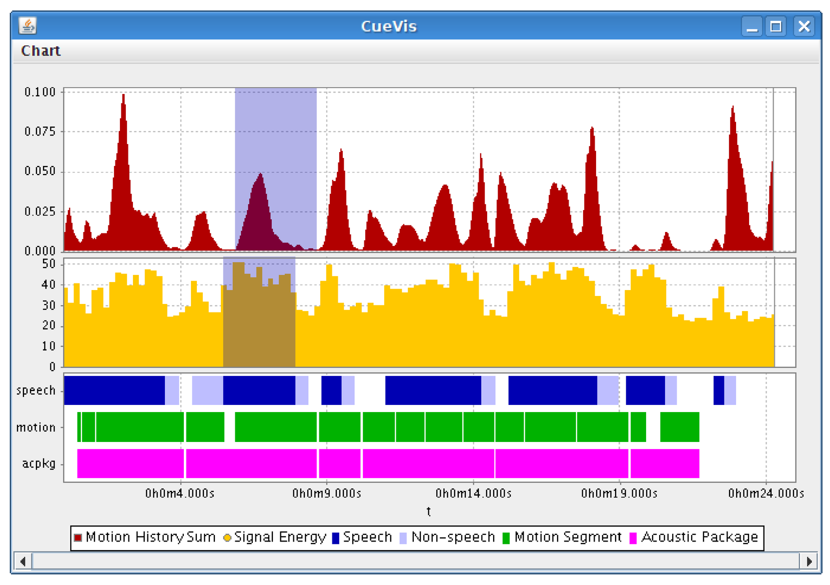

Figure: Cue visualization tool. The highlighted areas correspond to the motion peak and speech segment displayed in the inspection tool.

Evaluation

- Hypotheses
  - More acoustic packages in adult-child interaction than in adult-adult interaction.
  - Adult-adult interaction is less structured than adult-child interaction resulting in more motion peaks per acoustic package.
- We processed 11 videos with adults demonstrating the stacking of cups in both conditions.
- Results (AA vs. AC condition)
  - Significant difference in the number of acoustic packages: $t = 3.618$, $p = 0.005$
  - Significant difference in the ratio of motion peaks to acoustic packages: $t = 4.654$, $p = 0.001$

	Adult-Adult Interaction			Adult-Child Interaction		
Subj.	AP	M	M/AP	AP	M	M/AP
1	3	7	2.33	17	33	1.94
2	3	8	2.67	7	14	2.00
3	3	13	4.33	17	30	1.76
4	3	9	3.00	3	5	1.67
5	10	24	2.40	34	60	1.76
6	1	4	4.00	3	7	2.33
7	2	7	3.50	8	10	1.25
8	2	7	3.50	13	29	2.23
9	2	6	3.00	6	13	2.17
10	3	16	5.33	7	14	2.00
11	5	10	2.00	8	14	1.75
M	3.36	10.09	3.28	11.18	20.82	1.90
SD	2.42	5.70	0.99	8.99	16.10	0.30

Table: Counts of acoustic packages (AP) and motion peaks (M) on subjects in adult-adult interaction compared to the same adults interacting with children.

Conclusion

- We presented a first computational approach towards modeling acoustic packaging for human-robot interaction in a tutoring scenario.
- Our approach works on natural data.
- Our implementation follows a modular concept.
- Next steps targeting at:
  - Using acoustic packaging in designing a feedback of the iCub Robot.
  - Automatically deriving speech and action models from acoustic packages.

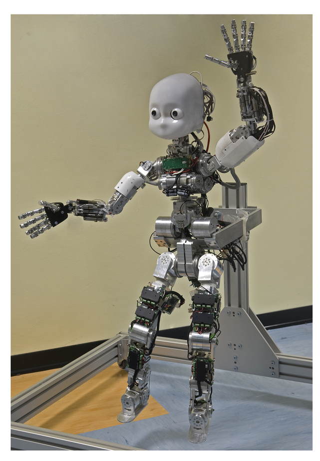

Figure: The iCub humanoid robot [robotcub.org].

References

- K. Hirsh-Pasek and R. M. Golinkoff, *The Origins of Grammar: Evidence from Early Language Comprehension*, The MIT Press, 1996.
- R. J. Brand and S. Tapscott, "Acoustic packaging of action sequences by infants," *Infancy*, vol. 11, no. 3, pp. 321–332, 2007.
- K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.
- J. Fritsch and S. Wrede, "An integration framework for developing interactive robots," 2007, pp. 291–305.