Universität Bielefeld

Acoustic Packaging and the Learning of Words

Lars Schillingmann[1] – Petra Wagner[2] – Christian Munier[3]
Britta Wrede[4] – Katharina Rohlfing[5]

[1 3 4]Applied Informatics Group, Faculty of Technology / [1 3 4]Research Institute for Cognition and Robotics
[2]Faculty of Linguistics and Literature / [5]Emergentist Semantics Group, CITEC
Bielefeld University, Germany

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].
- **Support for action and language 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.

System Overview

- Modular and decoupled approach
  - Modules communicate through a central memory: the Active Memory [2].
  - The Active Memory notifies components about event types they have subscribed to.
  - All modules are able to incrementally update their hypotheses based on the events they receive.

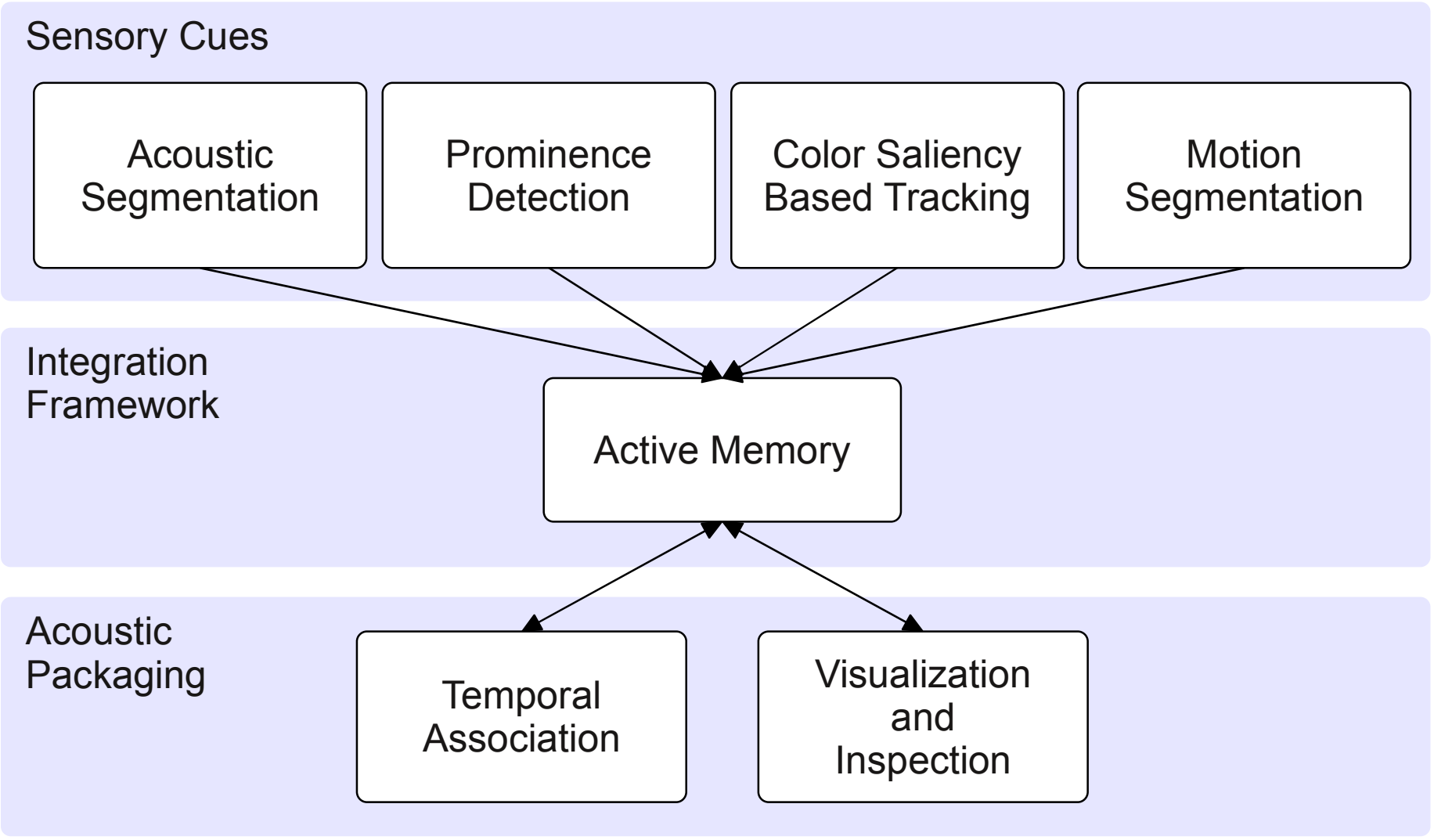


Figure: System overview with highlighted layers and their relation to the acoustic packaging system.

Multimodal Segmentation and Temporal Association [3]

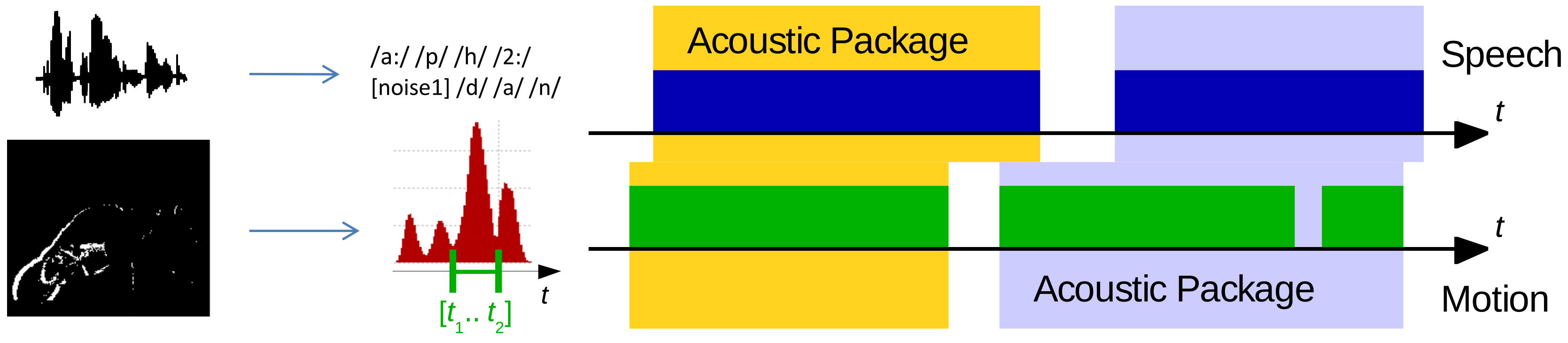


Figure: Segmentation of the visual and acoustic input cues into acoustic packages.

- Acoustic signal: Segmentation into speech and speech pauses
  - The ESMERALDA speech recognizer is used to detect voice activity more robustly than an approach that is solely based on signal energy.
- Visual signal: Segmentation into motion peaks
  - A peak ranges between two local minima in the amount of changed pixels in the visual signal.
  - The amount of changed pixels is calculated by summing up a motion history image at each time step.
- Temporal association: Overlapping speech and visual segments are associated to one acoustic package.

Detecting Moving Colored Objects

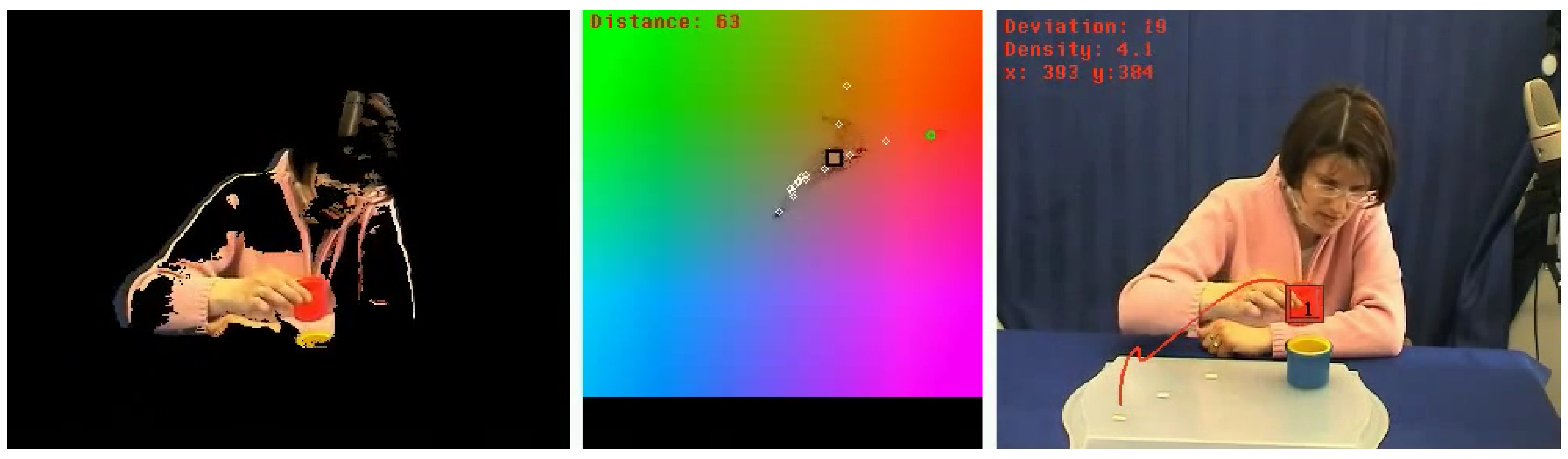


Figure: Processing steps of our color-saliency-based tracking module

1. Detecting changing regions using motion history images
2. Clustering in YUV color space and ranking according to color distance (U,V) to centroid of all clusters
3. Heuristical filtering (e.g. to detect clusters with uncovered background) and trajectory accumulation using Euclidean and color distance

Acoustic Prominence

- Idea: Relative ranking of syllables within an utterance [4].
- Syllable segmentation using the Mermelstein algorithm
- Spectral emphasis currently used as prominence feature
- Further prominence features: Nucleus duration, pitch movements, overall intensity
- Evaluation: 59.7% agreement to human rater (139 utterances, ∼4.45 words per utterance)

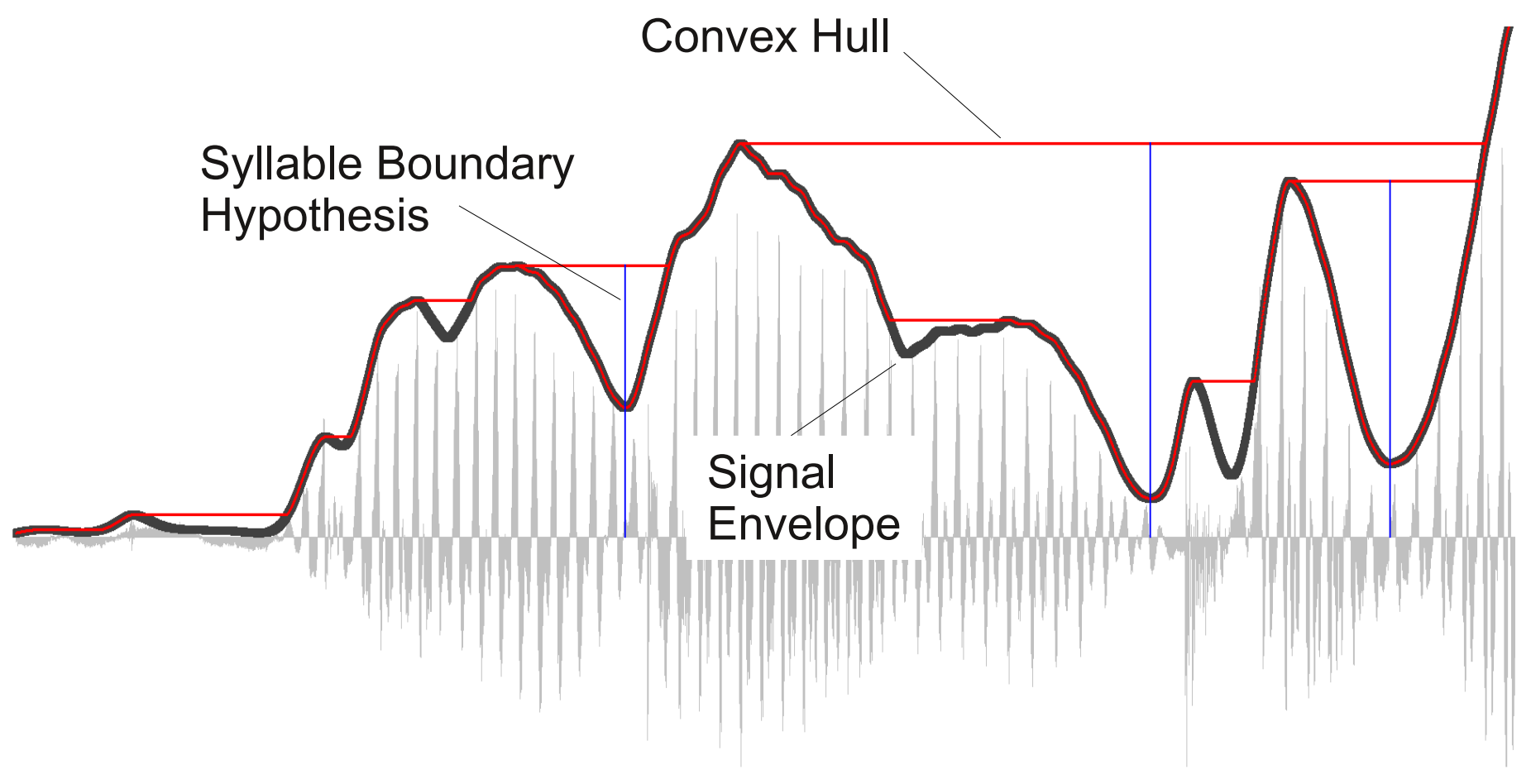
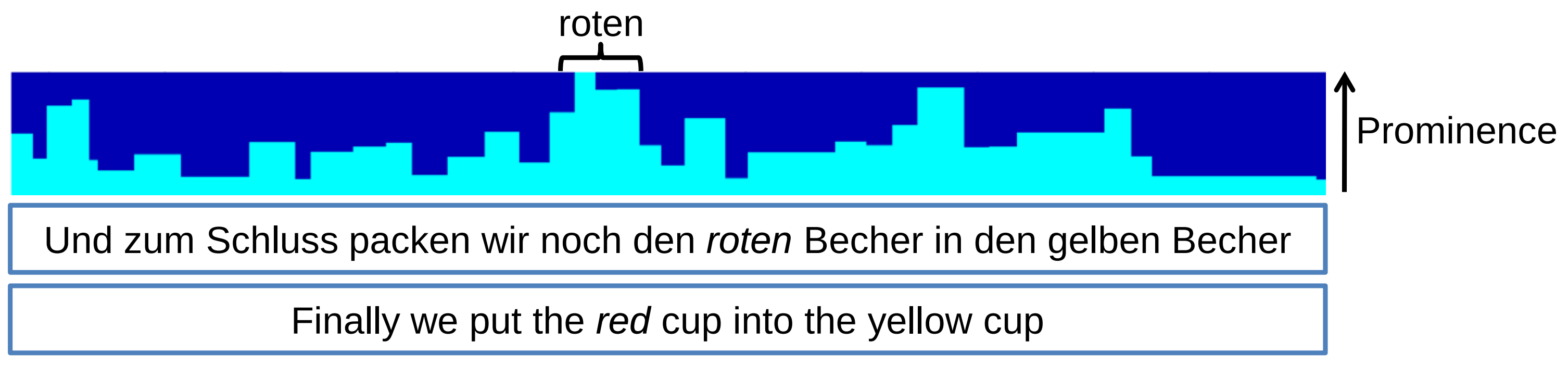


Figure: Mermelstein algorithm – Syllable boundaries are detected by approximating the signal's envelope with a convex hull.



Und zum Schluss packen wir noch den *roten* Becher in den gelben Becher

Finally we put the *red* cup into the yellow cup

Figure: Syllable segmentation and prominence ranking of an utterance.

Visualization and Inspection

- Analysis of temporal relations
- System inspection and debugging
- Rows (top–down)
  1. Motion activity
  2. x and y coordinates of object trajectories
  3. Acoustic signal energy
  4. Speech segmentation including prominence ranking
  5. Motion peaks segmented
  6. Acoustic packages formed

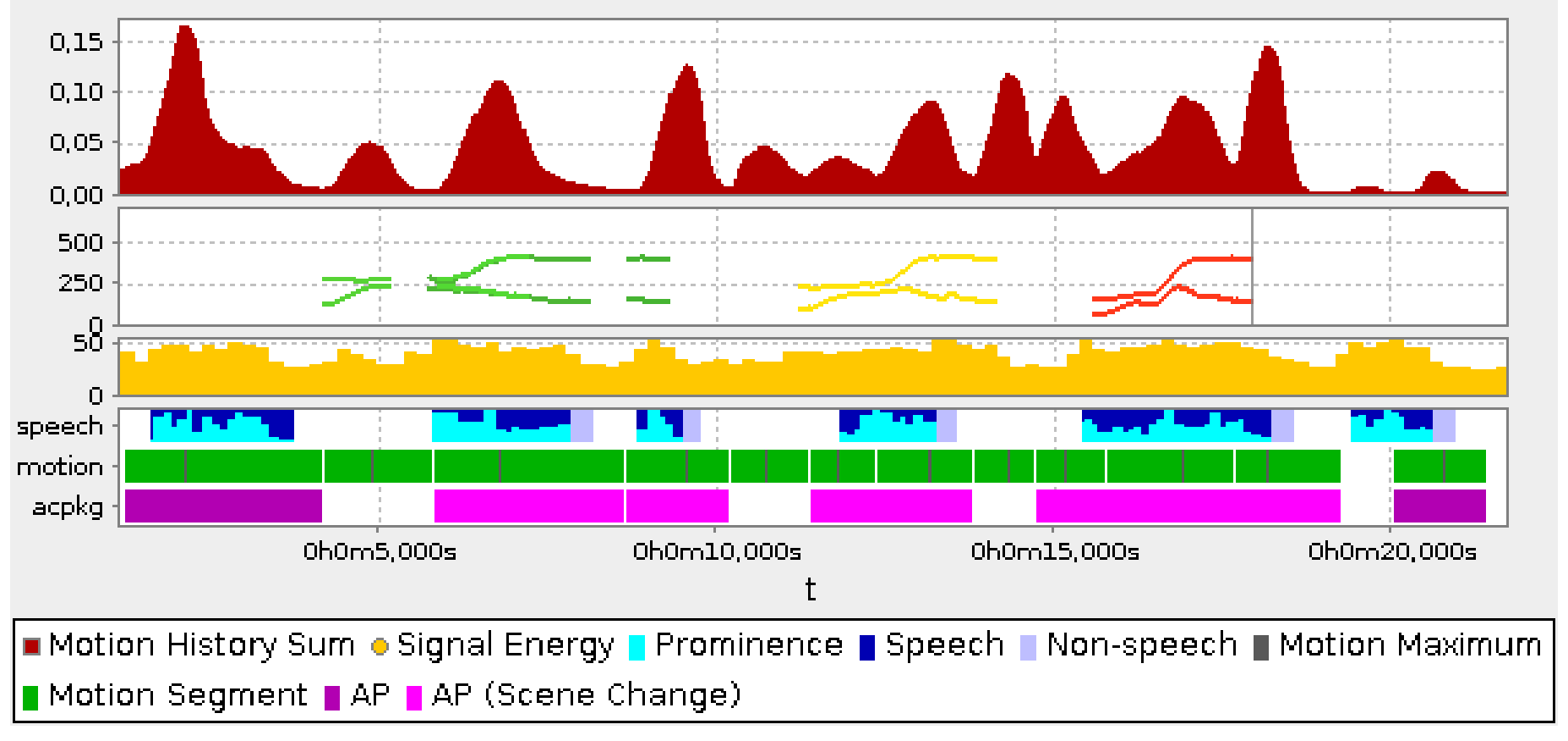


Figure: Visualization tool displaying cues extracted from a stacking cups task and their temporal development.

Acoustic Packages: Do Prominent Color Terms match Trajectory Colors?

- Visualization of color terms containing a prominent syllable in data of adults teaching children how to stack cups (see Figure).
- Color terms frequently match the object's trajectory color when highlighted
- However, many terms not referring to colors are also highlighted but filtered here.
- Tests on the iCub robot: Prominent syllables can be used to provide feedback that refers to semantically relevant parts of the utterance, such as color terms.

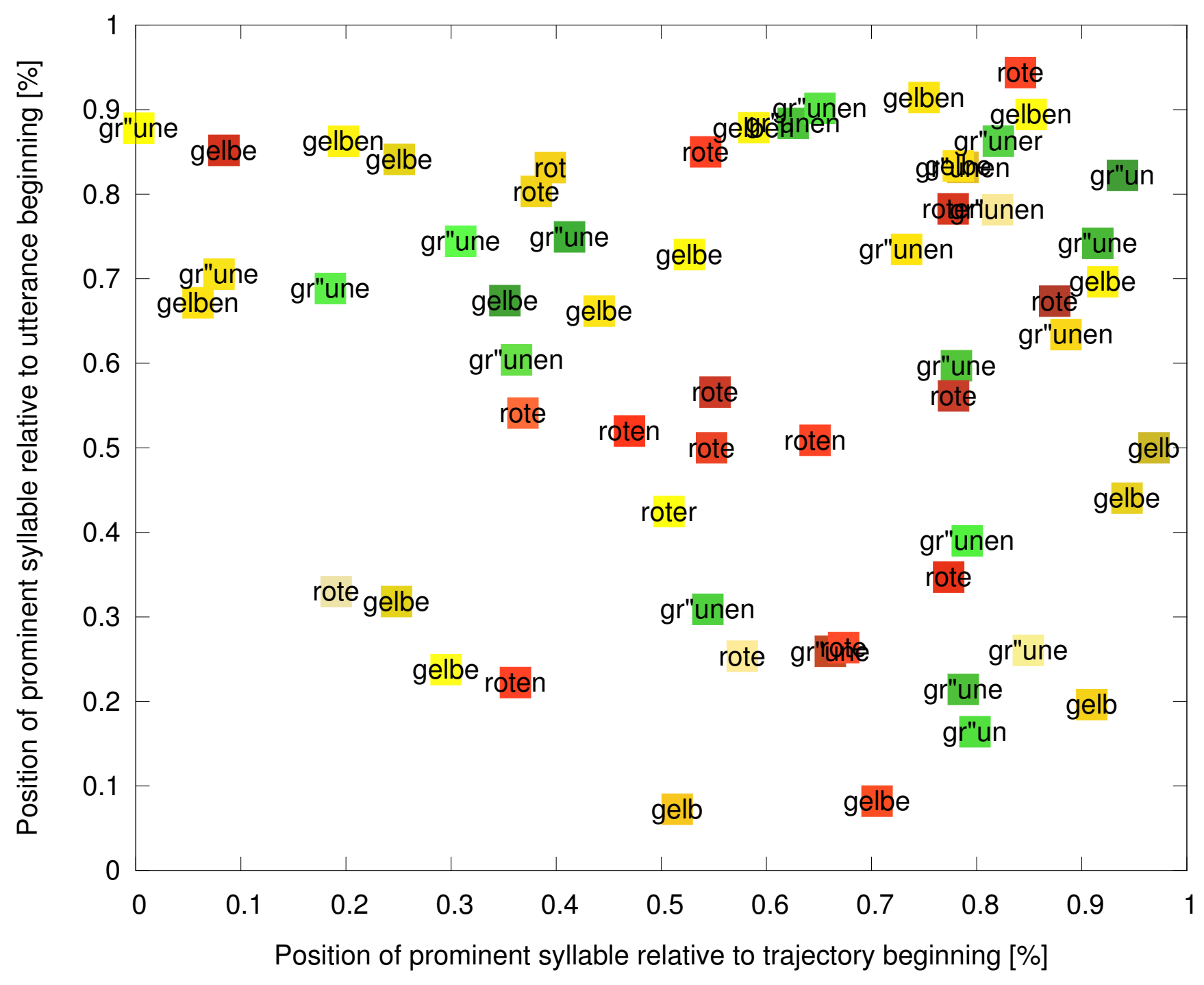


Figure: Prominent color terms and the trajectory colors detected.

Conclusion

- Color-saliency-based tracking and prominence detection were integrated into the acoustic packaging system.
- Acoustic packages simplify access to corresponding multimodal events at a given time.
- First steps towards word learning: The iCub robot can connect visual properties to highlighted linguistic units.

Figure: Demonstrating cup stacking to the iCub robot.

References

[1] K. Hirsh-Pasek and R. M. Golinkoff, *The Origins of Grammar: Evidence from Early Language Comprehension*. The MIT Press, 1996.

[2] J. Fritsch and S. Wrede, "An Integration Framework for Developing Interactive Robots," in *Software Engineering for Experimental Robotics*, D. Brugali, Ed. Springer, 2007, pp. 291–305.

[3] L. Schillingmann, B. Wrede, and K. J. Rohlfing, "A Computational Model of Acoustic Packaging," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 4, pp. 226–237, Dec. 2009.

[4] F. Tamburini and P. Wagner, "On automatic prominence detection for German," in *Interspeech 2007*, 2007, pp. 1809–1812.