Using Prominence Detection to Generate Acoustic Feedback in Tutoring Scenarios

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

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

{lschilli, bwrede, rohlfing, cmunier}@techfak.uni-bielefeld.de
petra.wagner@uni-bielefeld.de

Abstract

Robots interacting with humans need to understand actions and make use of language in social interactions. Research on infant development has shown that language helps the learner to structure visual observations of action. This acoustic information typically in the form of narration overlaps with action sequences and provides infants with a bottom-up guide to find structure within them. This concept has been introduced as acoustic packaging by Hirsh-Pasek and Golinkoff. We developed and integrated a prominence detection module in our acoustic packaging system to detect semantically relevant information linguistically highlighted by the tutor. Evaluation results on speech data from adult-infant interactions show a significant agreement with human raters. Furthermore a first approach based on acoustic packages which uses the prominence detection results to generate acoustic feedback is presented.

Index Terms: prominence, multimodal action segmentation, human robot interaction, feedback

1. Introduction

In tutoring scenarios with a human tutor and an infant learner the infant needs to be able to segment the continuous stream of multimodal information it perceives into meaningful units. In this context speech plays an important role for the segmentation process when the tutor is commenting his actions while demonstrating it: On utterance level, speech helps the learner attending to particular units of the action stream and connecting them. Within an utterance, emphasis helps the learner to identify relevant semantic information — for example the color of a described object, or the goal of an ongoing movement. If we build robotic systems able to learn actions, they could make usage of the speech-vision interaction in order to segment actions into meaningful parts in a way similar to infants. The idea that language helps infants to structure the action stream they perceive has been proposed and termed acoustic packaging in [1]. A computational model which is able to segment a continuous stream of speech and action demonstrations into acoustic packages has been proposed in [2]. Here, acoustic packages are designed as bottom-up units for further learning and feedback processes. Feedback is important to communicate to the tutor what the robot has understood from the tutor's action demonstrations. In this context it is important to further refine the action segmentation to identify highlighted parts. That way, the robot can report to the tutor that it learned the correct words or expressions for the action or term the tutor has focused on. For example, if the tutor showed a cup and focuses on the cup's color in the tutoring situation, s/he will probably emphasize the color term. By repeating this (emphasized) color term, the robot shows its understanding of the dialogue's essence.

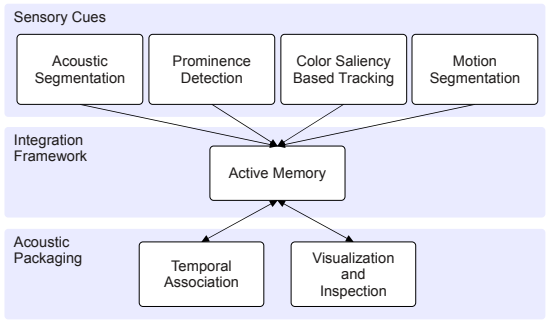

Figure 1: System overview with highlighted layers and their relation to the acoustic packaging system.

In this paper we will focus on identifying these parts in the acoustic modality. We will utilize perceptual prominence [3] to detect highlighted syllables in action presentations. In the next sections we will describe our recent work: the implementation of prominence detection and its integration into our architecture for multimodal action segmentation.

2. Acoustic Packaging

The development of robots, which are able to interact with humans in tutoring situations, requires methods to segment actions into meaningful parts. In [2] we describe a system where we transferred the concept of acoustic packaging from developmental research to fulfill two important tasks in human-robot tutoring situations. The first task is to deliver bottom-up segmentation hypotheses about the action presented. The second task is to form early learning units containing multimodal information. These units can further be processed by other modules that infer models about the actions currently presented.

The acoustic packaging system has to fulfill three main requirements. Since the system integrates visual and acoustic cues a *temporal segmentation* for both modalities is required. A second problem is the *temporal synchronization* of these sensory

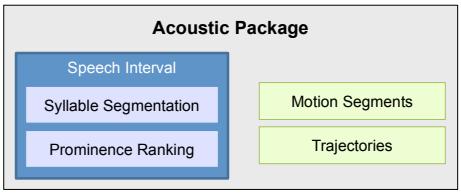

Figure 2: Overview of the sensory cues which are associated to an acoustic package

cues. Hypotheses from audio and vision processing are typically generated neither at the same time nor in the same rate. The system considers temporal synchrony as an amodal cue, which provides information about what segments should be packaged. Since a socially interactive robot should give feedback during tutoring, the system has to be usable *online* and able to cope with updating hypotheses.

The acoustic packages generated by this system are further processed by a feedback module which provides basic feedback based on prominent syllables. In the following we will give a summary of the different processing modules in the acoustic packaging system including the prominence detection module. In the last section the feedback module will be described.

System Overview Our system for acoustic packaging consists of six modules (see Figure 1). These modules communicate events through a central memory, the so called active memory [4]. The active memory notifies components about event types they have subscribed to and is able to store these events persistently. It is an integration framework which supports a decoupled design of the participating modules facilitating integration of further processing modules.

Acoustic Segmentation The audio signal is segmented using the ESMERALDA speech recognizer [5], which is configured to use an acoustic model for monophoneme recognition. A continuous chain of phoneme hypotheses generated by the speech recognizer is considered a speech segment. This method provides a more robust approach than a simple VAD approach, such as detecting speech based on signal energy. This applies especially in tutoring situations with noisy acoustic conditions. Our speech recognizer inserts those phoneme hypotheses as well as the corresponding audio signal into the active memory. As the recognition process is incremental during processing of an utterance the hypotheses are continuously updated.

Prominence Detection We understand perceptual prominence of linguistic units as the unit's degree of standing out of its environment [3]. This results in two main requirements for this module which automatically detects perceptual prominent units. First the speech stream has to be segmented into linguistic units, which in our case are syllables. This type of units is typically used in prominence detection methods and furthermore has the advantage that speech can be segmented into syllables without using models that require a known lexicon. The second step is to rate these linguistic units according to the acoustic parameters which correlate to the perceived prominence. Our implementation and evaluation of the prominence detection module will be described in more detail in Section 3. If a new utterance hypothesis is completed, the prominence detection module retrieves the acoustic signal from active memory and performs

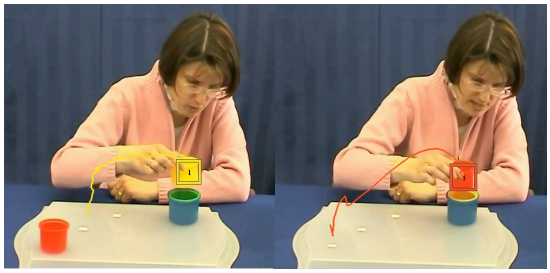

Figure 3: Two examples of tracking results using the color saliency based tracking module. The images show a test subject demonstrating cup stacking to an infant. The color property of the trajectory is automatically determined from the salient regions tracked.

the steps above. The result is a syllable segmentation which includes a prominence rating for each syllable. The utterance hypothesis is extended with this information and made available to other modules by inserting the updated hypothesis into the active memory.

Visual Action Segmentation The visual signal is segmented into motion peaks where each peak ranges between two local minima in the amount of change in the visual signal. For example, if someone shows a cup, there is typically a motion minimum at the point where the cup is held still or slowed down for a short moment. When the cup is accelerated again on its way to be put on a table, a local maximum in the amount of motion can be observed. Another local minimum occurs when the cup is eventually put on the table. This observation is the motivation for our heuristic approach to segment actions into motion peaks. This segmentation into motion peaks is technically realized by an approach based on motion history images [6]. The amount of motion is calculated per frame by summing up the motion history image. In the amount of motion local minima are detected with the help of a sliding window that is updated at each time step.

Color Saliency Based Tracking The visual action segmentation can provide a temporal segmentation of the video signal but cannot deliver detailed spatial information and local visual features about moving objects in the tutoring situations. Our approach is based on the assumption that during action demonstrations the objects are typically moved. Furthermore we assume uniformly colored objects. A short summary of our method is presented in the following: The visual signal is masked using a motion history image to focus on the changing parts in the visual signal. The pixels of the changing regions are clustered in the YUV color space using UV coordinates for the distance function. The clusters are ranked according to their distance to the center of mass of all clusters. The top ranked clusters are considered salient. Several heuristics are applied to filter out e.g. background which is uncovered. The top ranked clusters are tracked over time based on spatial and color distance. The top ranked trajectory forms the motion hypothesis of the object presented by the tutor (see Figure 3).

Temporal Association The motion peaks, trajectory hypotheses, and speech segments need to be temporally associated in order to form acoustic packages. Our temporal association module subscribes to events on the active memory and maintains

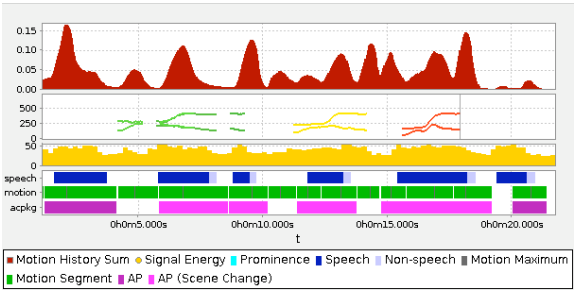

Figure 4: Cue visualization tool showing the segmentation and association of speech, prominence, motionpeaks, and trajectories to acoustic packages.

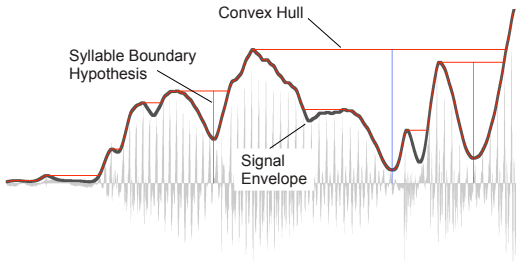

Figure 5: Visualization of the Mermelstein convex hull based syllable segmentation algorithm. The convex hull is drawn at multiple iterations to visualize its approximation of the energy envelope.

a timeline for different types of time intervals. In our current version of the system motion peaks, trajectories, and speech segments (including prominence detection results) are processed. When a new event arrives, the segment is aligned to the timeline. In the next step, the temporal relations to the segments on the other timeline are calculated for which a subset of the relations defined in [7] is used. When overlapping speech and motion segments are found on the timelines acoustic packages are created. In the case that motion segments overlap with two different speech segments, the one with the larger overlap is chosen. Trajectory segments are associated with acoustic packages using the same strategy: If they overlap with a speech segment they are associated to the according package. If hypotheses from the signal processing modules are updated (e.g. a speech segment is extended), the corresponding acoustic package is updated as well. Figure 2 gives an overview of the different cues associated to acoustic packages.

Visualization and Inspection Since temporal synchrony is one important cue for this system, tools are needed that analyze the acoustic packaging process and the temporal relations of the involved sensory cues. Figure 4 shows our visualization tool, monitoring events which are communicated to the active memory by other processing modules. The first plot displays the amount of motion over time. The second row shows the x and y coordinates of the trajectories which have been tracked by the color saliency module. The third row shows the signal energy that gives an estimate about speech activity. The fourth row visualizes the hypotheses as time intervals coming from the acoustic segmentation, the visual action segmentation, and the temporal association module. More specifically, the first line displays the speech recognition results: The lighter areas mark non-speech hypotheses like for example noise. Within each speech segment the syllable segmentation and prominence rating is displayed as bars. The highest bar is the most prominent syllable. The second line displays the temporal extensions of the motion peaks. The third line visualizes the results of the acoustic packaging module. Since the case is possible that under certain conditions the temporal extensions of two neighboring acoustic packages overlap, only the range of motion peaks (which have been associated to one acoustic package) is currently visualized.

Feedback Based on Prominent Syllables Acoustic packages contain temporally overlapping intervals from different modalities. Thus, acoustic packages simplify the access to corresponding multimodal events at a time. In our scenario, a human tutor sits in front of a robot and demonstrates cup stacking to the system. A typical acoustic package contains an utterance hypothesis (including prominence ratings) and a trajectory hypothesis. The idea of our feedback module is to make use of the trajectory color information and the prominent syllables in the acoustic packages in order to associate semantically relevant syllables from speech with properties of the object presented. During action demonstrations with the tutor explaining his actions the acoustic packages are clustered using the color feature of the object trajectories. When the tutor tries to evaluate what the system has learned by only showing the cups but not explaining his actions the feedback module complements the speech modality by replaying the most prominent syllable from a package belonging to a cluster with a similar color. A neighborhood of two syllables is included in replaying as a heuristic to ensure capturing a full word and to compensate for possible oversegmentation effects.

3. Perceptual Prominence

As described in Section 2, our module segments the speech stream into syllables. Subsequently, each syllable is rated according to the acoustic parameters correlating with perceived prominence. These acoustic parameters have to be chosen carefully according to their robustness in noisy acoustic environments. Furthermore, their implementation needs to be fast enough to be applied in scenarios which require online feedback.

Syllable Segmentation A modified version of the Mermelstein algorithm [8] is used to segment utterances into syllables. In a first step the signal is filtered using an equal loudness filter [9]. The filtered signal is further bandpass filtered using a 4th order Butterworth bandpass filter with a lower cut-off frequencies at 500 Hz and 4000 Hz. Then, the signal is full wave rectified and low-pass filtered with a second order Butterworth filter at 40 Hz to obtain an estimation of the signal's envelope. The basic idea of the Mermelstein algorithm is to detect minima in the signal's energy envelope. The locations of these minima are the desired syllable boundaries. The minima detection is described in the following: The signal's envelope is approximated using a convex hull. A syllable boundary is identified at the maximum difference between the convex hull and the signal's envelope (see Figure 5). The algorithm is carried out recursively for the intervals left and right to the syllable boundary. The recursion is terminated if the maximum distance drops below a certain threshold or the interval between two boundaries falls below a minimal length. The general idea behind this approach is to prioritize the most significant minima in the signal's envelope.

Prominence Rating The algorithm described in [3] uses the following set of acoustic parameters which contribute to the perceived prominence in German: nucleus duration, spectral emphasis, pitch movements, overall intensity. These acoustic parameters are weighted and combined to the resulting prominence measurement. In [3] the best results are achieved using high weighting factors for spectral emphasis and nucleus duration features. To simplify the algorithm we focus here on the most relevant and robust parameters for acoustic prominence as an initial step: Spectral emphasis is used to rate the syllable segments. The syllable segment with the highest spectral emphasis rating is considered the most prominent syllable in the utterance. The spectral emphasis feature is calculated by bandpass filtering the signal from with a 4th order Butterworth filter in the band 500 Hz to 4000 Hz. Then, RMS energy is computed for each syllable segment and normalized per utterance. Since a nucleus duration feature would depend on the accuracy of the syllable segmentation it is also left out in the first version of this approach for reasons of robustness.

Evaluation We evaluated both our syllable segmentation approach and the prominence rating method. Syllable segmentation was evaluated on a subset of the Verbmobil corpus [10], since an accurate syllable segmentation is available. The subset consists of 2,000 randomly selected utterances containing 68,276 syllables in total. A syllable boundary is considered a match if a boundary hypothesis is within 50ms distance. Table 1 shows results with balanced insertion and deletion rates.

Matches	Deletions	Insertions
68.65%	31.35%	31.35%

Table 1: Evaluation results of our syllable detection method on utterances from the Verbmobil corpus.

The prominence rating algorithm has been evaluated on a corpus with adult-infant interactions [11]. For the evaluation we used a subset where adults explain children how to stack cups. The acoustic channel has been recorded from a distant microphone and thus contains environmental noise e.g. from the cup stacking task and in some cases from the child. Word boundaries were automatically determined from a transcription by performing a forced alignment. A human annotator has marked the most prominent word in each utterance. If the center of the syllable with the highest prominence ranking lies within the word boundaries, a match is counted. Utterances with very bad acoustic conditions where even the forced alignment failed where not taken into account. 139 utterances have been used in the evaluation. The results are presented in Table 2.

Matches	Utterances	*Words
59.71%	139	4.45

Table 2: Evaluation results of prominence detection approaches on utterances from adult infant interactions. The results are 2.7 times better than chance. (*Average number of words per utterance)

4. Discussion and Conclusion

In this paper we described a prominence detection module and its integration in our acoustic packaging system where it detects semantically relevant information linguistically highlighted by a tutor. Evaluation results on speech data from adult-infant interactions show a 59.7% agreement with human raters. Furthermore a first approach based on acoustic packages which uses the prominence detection results to generate acoustic feedback was presented. Although the prominence module's agreement with a human rater is not perfect, the method works in the more difficult acoustic conditions of tutoring scenarios. Furthermore, the method definitely works considerably better than chance. Our method thus implements a successful strategy of human infant learners to extrapolate word-meaning pairs out of running discourse. By using more complex acoustic features the results could possibly be improved. Especially including nucleus duration would likely lead to an improvement as long as it is estimated robustly.

First tests of the feedback module on the iCub robot showed that our prominence detection module is able to facilitate feedback which refers to semantically relevant parts of the utterance. However, to close the loop between tutor and robot strategies for handling corrections or other types of feedback regarding the quality of the acoustic packages have to be implemented. Such methods would allow for developing the system to adapt to the tutor and keep only packages which maintain information also considered as relevant by the tutor.

5. Acknowledgements

The authors gratefully acknowledge the financial support from the FP7 European Project ITALK (ICT-214668).

6. References

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

[2] 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.

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

[4] 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.

[5] G. A. Fink, "Developing HMM-Based Recognizers with ESMERALDA." Springer-Verlag, 1999, pp. 229–234.

[6] J. W. Davis and A. F. Bobick, "The Representation and Recognition of Human Movement Using Temporal Templates." IEEE Computer Society, 1997.

[7] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983.

[8] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[9] D. Robinson, "Replay Gain - A proposed standard." [Online]. Available: http://replaygain.hydrogenaudio.org/

[10] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon, "Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL – 3.0," Sep. 1994.

[11] 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.