Chapter 61
Making Use of Multi-Modal Synchrony:
A Model of Acoustic Packaging to Tie Words to Actions

Britta Wrede
Bielefeld University, Germany

Lars Schillingmann
Bielefeld University, Germany

Katharina J. Rohlfing
Bielefeld University, Germany

ABSTRACT

If they are to learn and interact with humans, robots need to understand actions and make use of language in social interactions. Hirsh-Pasek and Golinkoff (1996) have emphasized the use of language to learn actions when introducing the idea of acoustic packaging in human development. This idea suggests that acoustic information, typically in the form of narration, overlaps with action sequences, thereby providing infants with a bottom-up guide to attend to relevant parts and to find structure within them. The authors developed a computational model of the multimodal interplay of action and language in tutoring situations. This chapter presents the results of applying this model to multimodal parent-infant interaction data. Results are twofold and indicate that (a) infant-directed interaction is more structured than adult-directed interaction in that it contains more packages, and these packages have fewer motion segments; and (b) the synchronous structure within infant-directed packages contains redundant information making it possible to solve the reference problem when tying color adjectives to a moving object.

DOI: 10.4018/978-1-4666-4607-0.ch061

INTRODUCTION

Learning robots are the holy grail of robotics research. However, research in this area tends to focus on algorithms that function autonomously without the influence of, for example, a teacher. Action imitation learning approaches, for instance, explicitly exclude the interactive situation by recording a teacher's demonstration with the starting point being signaled by the robot. Note, however, that speech learning already reveals some unimodal approaches that take the tutor's feedback into account (e.g., Sato, Ze, & Dijk, this volume; also Arsenio, this volume). Nonetheless, in general, these approaches take no account of the speech input at all. This is in stark contrast to recent findings in developmental linguistics that point to the importance of not only social feedback from the tutor but also information arising from the situation itself for understanding and learning about objects and actions along with their meaning. Yet, analyzing social feedback and situational aspects is a very difficult algorithmic problem. From this perspective, acoustic packaging provides an intriguing mechanism with which to segment multimodal data streams into meaningful units based on the assumption that speech provides meaningful action boundaries.

In developmental research, Hirsh-Pasek and Golinkoff (1996) have proposed acoustic packaging as one possible way of performing bottom-up action segmentation. It has been suggested that this form of bootstrapping guides children toward the hierarchically organized action structure found in adults (Zacks & Tversky, 2001). Hirsh-Pasek and Golinkoff (1996) distinguished between a minimal and a maximal role of acoustic packages. In the minimal role, acoustic packages are formed when an acoustic segment is repeated in synchrony with an action event. In the maximal role, in contrast, acoustic packaging can fuse separate events into meaningful macroevents. In this manner, speech may help infants to understand that certain components of a diapering routine go together, and that they are separable from the following "clean-up" event. Indeed, acoustic packaging has been shown to influence infants' interpretation of demonstrated actions, because they tend to tie together actions that are accompanied by speech while considering action parts not accompanied by speech as not belonging to the action (Brand & Tapscott, 2007; Stouten, Demuynck, & Van Hamme, 2007).

Acoustic packaging is also important from a robotics perspective: To enable a robot to learn actions, it has to figure out which part of the demonstration belongs to the meaningful part of an action and which part it should ignore. Similarly, for a speech learning system to be able to relate meaning to acoustic patterns, the system needs not only to segment speech into recurring patterns such as phones or syllables (Brandl, 2009) but also to tie it to meaningfully segmented events in the real world. However, current robotics approaches to multimodal speech learning tend to rely on artificially created data that carefully exclude the multimodal segmentation problem (Van Hamme, 2007).

In addition, the concept of acoustic packaging complements research in action segmentation in two important ways: First, it provides a developmental perspective on action understanding, showing how meaningful units that are crucial for action perception and action memory can be learned in adults. Second, it adds acoustic signals to the mostly visual features that have been proposed as contributing to bottom-up processing for meaningful action parsing in adults (Zacks & Swallow, 2007) and in infants (Baldwin, Baird, Saylor, & Clark, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007). Acoustic packaging therefore integrates both features for visual and acoustic segmentation, while regarding their modality-specific properties. In spite of its transient nature, acoustic information impacts young infants' attentional focus more than concurrent visual information (Robinson & Sloutsky, 2004). This aspect is reflected in our model (see Figure 1).