Acoustic Packaging: Maternal Speech and Action Synchrony

Meredith Meyer, Bridgette Hard, Rebecca J. Brand, Molly McGarvey, and Dare A. Baldwin

***Abstract*— The current study addressed the degree to which maternal speech and action are synchronous in interactions with infants. English-speaking mothers demonstrated the function of two toys, stacking rings and nesting cups, to younger infants (6-9.5 mo.) and older infants (9.5-13 mo.). Action and speech units were identified, and speech units were coded as being ongoing action descriptions or non-action descriptions (examples of non-action descriptions include attention-getting utterances such as 'Look!' or statements of action completion such as 'Yay, we did it!'). Descriptions of ongoing actions were found to be more synchronous with the actions themselves in comparison to other types of utterances, suggesting that 1) mothers align speech and action to provide synchronous "acoustic packaging" during action demonstrations and 2) mothers selectively pair utterances directly related to actions with the action units themselves rather than simply aligning speech in general with actions. Our results complement past studies of acoustic packaging in two ways. First, we provide a quantitative temporal measure of the degree to which speech and action onsets and offsets are aligned. Second, we offer a semantically-based analysis of the phenomenon, which we argue may be meaningful to infants known to process global semantic messages in infant-directed speech. In support of this possibility, we determined that adults were capable of classifying low-pass filtered action- and non-action-describing utterances at rates above chance.**

***Index Terms*—Action processing, pediatrics, speech processing, multimodal communication**

I. INTRODUCTION

ON a daily basis, we witness hundreds if not thousands of intentional human actions being performed in the world. In order to make inferences about their causes and the goals that motivate them, one first step is to segment the action stream into units. For instance, imagine trying to understand the action stream encountered during a typical meal preparation. Recognizing the existence of action units such as "cutting a vegetable" or "washing a dish" helps us make sense of the busy flow of motion; we are better able to analyze the action stream on the basis of inferences about the actor's goals. How is it that we come to be able to segment the action stream? Segmentation faces a number of in-principle problems that are relevant to this question; action is often continuous, without pauses reliably marking their onsets and offsets. Further, actions may overlap, and parts of the action stream are often obscured or may be difficult to see [1], [2].

Despite such complexity, action segmentation appears to be relatively effortless, automatic, and spontaneous [3]-[5]. People typically report onsets and offsets of actions that appear to coincide with the initiation and completion of goals, and studies with adults find remarkable consistency in where people report the location of these action breakpoints [1], [4], [6]. Upon reflection on one's own phenomenological experience of action, it seems clear that top-down inferences generated by our own knowledge of goals and intentions aids in this segmentation process. For example, imagine witnessing a waiter at a restaurant bringing food to a table. Familiarity with the waiter's intention to serve the correct plates to the diners, recognition of actions associated with serving food, and an understanding of the need to balance heavy plates and full drinks all aid us in recognizing where actions units begin and end [7].

Top-down processes involved in action segmentation are unlikely to account for the entire story, however. For example, Baldwin and colleagues found that infants parse dynamic human action in units corresponding to the completion and initiation of intentions. Ten-month-old infants who had first been familiarized to a simple action stream (a woman dropping a towel and bending down to pick it up) responded with increased attention when pauses were inserted within action units (i.e., in the middle of bending down) as opposed to when pauses fell at action boundaries (i.e., at the moment the towel was grasped) [8]. In another study, infants as young as nine months also preferred to watch a display of dynamic human action where tones matched action boundaries as opposed to where they did not coincide with boundaries [9].

Given that infants probably lack sophisticated theory of mind skills and knowledge of the intentions that motivate even ordinary, everyday adult actions, the fact that infants appear to process the action stream by segmenting it into goal-relevant units is remarkable. The fact that this ability appears to be in place well in advance of extensive top-down knowledge about goal-directed action implies that bottom-up mechanisms are likely supporting such action segmentation. One possible mechanism might involve infants' sensitivity to acoustic packaging, the phenomenon under investigation in the current study.

Manuscript received August 20, 2010

Meredith Meyer was with the University of Oregon, Eugene, OR 97403 USA. She is now with the Department of Psychology, University of Michigan, Ann Arbor, MI 48103 USA (734-615-0575; email: mermeyer@umich.edu).
Bridgette Hard is with Stanford University, Palo Alto, CA 94305 USA. (email: brimart@stanford.edu).
Rebecca J. Brand and Molly McGarvey are with Villanova University, Villanova, PA 19085 USA. (email:rebecca.brand@villanova.edu/ mmcgar08@villanova.edu).
Dare Baldwin is with the University of Oregon, Eugene, OR 97403 USA. (email: baldwin@uoregon.edu).

Acoustic packaging is a form of multimodal input in which parents or other caregivers align verbal utterances with events. For instance, a mother dressing her infant might narrate such events as putting on a shirt or fastening the legs of a onesie by saying, "Put your shirt on" or "Let's get all the snaps, one, two, three", aligning her utterances with the associated actions. Although originally proposed as a way of helping children to discover the structure of *language* by highlighting relevant units such as phrases or clauses [10, 11], it is also the case that the multimodal information provided by such packaging could help infants detect the relevant units in a continuous action stream. Specifically, if infants were sensitive to such alignment, they might be better able to detect the onsets and offsets of actions via attention to the onsets or offsets of phrases or words. But to what degree is acoustic packaging actually a normal or typical part of the infant's life? Very little research has addressed this topic (though see work by Schillingmann and colleagues [12] and Rolf and colleagues [13]). In the current paper, we present an observational study of maternal acoustic packaging, reporting on the degree to which maternal speech and action are synchronous in interactions with preverbal infants. We also examine the semantic content of maternal utterances, asking whether utterances directly relevant to actions themselves may be preferentially used to structure synchronous interactions with infants.

II. Modifications to Infant-Directed Communication

In some ways, acoustic packaging can be construed as an extension to a phenomenon that has already been extensively studied, namely maternal modifications to speech. Ample evidence indicates that mothers and other caregivers typically modify their utterances in the presence of infants. Adults in many cultures tend to use shorter, simpler utterances and to exaggerate intonation in ways that express certain global emotional messages [14]-[16]. For example, mothers speaking in a number of different languages (e.g., American English, Japanese, Hausa) tend to use a highly similar set of distinctive intonational patterns to get their infants' attention (increased pitch and pitch excursions), soothe their infants (decreased pitch and more fluid utterances), or express prohibition or disapproval (sharp, stacatto bursts) (for a review, see [15]).

Through these modifications, mothers are potentially providing infants with access to meaning well before infants have come to be able to comprehend the meanings of the individual words being uttered. The prosodic features may themselves be enough to convey overall content; adults are capable of classifying infant-directed utterances into global semantic categories even when the utterances are not in the adult subjects' native language or when the utterances are low-pass filtered, obscuring articulatory features required for individual word identification [16], [17]. As well, infants themselves also appear to recognize the general valence of such utterances; for example, infants smile more when hearing utterances featuring approval contours in comparison to prohibitive utterances, even when the language they are hearing is not their native language [18].

Infant-directed speech also is believed to support language learning itself, ranging from discovery of statistical structure of syllables in continuous speech [19], to word-learning [20], to phrase recognition [21]. Note that these accounts of specific linguistic facilitative effects and the accounts focusing on global semantic function are not mutually exclusive. Instead, it is suggested that the more global emotional semantic functions play a role earlier in development, with specific linguistic effects becoming more prominent as the child starts to understand and use language [22].

Researchers have also observed similar modifications to infant-directed *action*. In the first study to demonstrate such infant-specific modifications to action, Brand and colleagues videotaped mothers' demonstration of a set of toys to either their 6- to 13-month-old infants or a familiar adult family member or friend [23]. Analysis of the videos revealed that mothers' actions with infants featured increased enthusiasm, repetition, simplification, and interactiveness relative to their interactions with an adult partner. Mothers also held the toys closer to their infants and broadened the range of their motions (see also [24] for similar results). Based on the striking similarities seen in infant-directed action relative to infant-directed speech, the authors dubbed this phenomenon *motionese* (derived from the word *motherese*, another word used for infant-directed speech). Indeed, Brand and colleagues suggested that this suite of modifications may be analogous to many of the adjustments seen in infant-directed speech; for instance, the increased pitch and pitch ranges seen in much of speech directed to infants may be expressed as increased proximity and broadening of motion in action.

Rohlfing and colleagues have also undertaken investigations of infant-directed action, focusing on the implications that this phenomenon might have for robot learning. More specifically, they note that robots and infants may face a set of similar problems when attempting to learn from and imitate action. Neither robots nor infants come equipped with sophisticated top-down knowledge that can be applied to detecting goal-directed units. Thus, one set of problems relates to "what to imitate"; in other words, both the robot and the infant need to determine which observed actions are relevant units with respect to the goal state of the actor. A second set of problems relates to "how to imitate"; robots and infants are also faced with the task of translating observed action into their own efficacious actions that bring about their intended outcomes in the world [25].

Rohlfing and colleagues suggest that part of the solution may lie in the modifications typical of infant-directed action; modifying the action stream by supplying pauses between units and exaggerating motion contours may direct the learner (either human or machine) to what is relevant by reducing the complexity of the signal. To investigate the nature of infant-directed action on a fine-grained and objective level, they have employed quantitative analyses of infant-directed action, using a 3-D body tracking system originally designed for human-robotic interaction [26]. Among their findings: Mothers and fathers tended to execute less "rounded" motions with their infant partners in comparison to their adult partners (e.g., a lift of a cup to a location would involve a pronounced vertical lift followed by a similarly exaggerated horizontal motion, rather than simply moving the cup in a smooth arc), and parents also paused more between the individual motions [27].

III. Multimodal Modifications: Speech and Action Synchrony

The preceding describes findings demonstrating that parents appear to modify both speech and action when interacting with their children. Another body of research has examined how parents structure communication with infants with respect to the *integration* of speech and action, examining how modifications to the two modalities together might scaffold infants' learning. Researchers interested in multimodal learning have noted that infants are capable of detecting redundancy from a young age; for instance, 4-month-old infants presented with auditory speech syllables prefer to look at mouth movements that match these sounds [28], and 5-month-old infants prefer to watch displays of intensity change when bimodal information matches (e.g., a mouth opens as sound amplitude increases) [29].

These and other related findings have given rise to the Intersensory Redundancy Hypothesis [e.g., 30], which in part states that such multimodal redundancy functions to recruit infants' attention, thereby aiding infants in extracting certain regularities from their perceptual environment. Indeed, evidence suggests that redundancy can have facilitative effects in infant learning; for example, 5-month-old infants presented with an animation of a hammer striking a surface later recognized changes to the rhythm only if they had both seen and heard the event, but not if they only perceived the event in one modality (i.e., saw the hammer without sound, or heard the hammer without seeing it). Notably, it was not just receiving information in both modalities that enabled this learning; only when the hammer strikes were synchronous with respect to both the visual and auditory syimulus were infants capable of learning the rhythm and detecting changes [31].

Of particular importance to the current study, synchrony between action and speech is one possible means of directing infants to the most relevant parts of the signal. Gogate and colleagues have investigated synchrony's role in word learning in a series of studies. In experimental studies, Gogate and Bahrick found that young infants (7-8 months of age) were only able to learn syllable-to-object mappings when the syllables were presented synchronous with object movement [32]. Complementing these findings, Gogate, Bahrick, and Watson demonstrated that parents are especially likely to align speech and gesture in a naming context (e.g., movement of a novel toy at the same time as uttering the label). They also discovered that this phenomenon undergoes change across infant development; the most synchrony was seen at the age at which the authors argue such multimodal cueing would be most useful to infants (pre-lexical, 5-8 months). By the time infants were older and could be assumed capable of exploiting other cues (such as understanding of referential intent, e.g., [33]), the degree to which mothers aligned their naming utterances and gestures decreased [34].

Gogate and colleagues' work has primarily focused on gesture-speech synchrony's role in language learning; however, as noted earlier, the facilitative attention and learning effects may work for action processing as well, with synchrony serving to highlight goal-relevant *action* units. Thus far, to our knowledge only one experimental study has been conducted with this effect in mind. Specifically, Brand and Tapscott investigated how acoustic packaging might impact infants' segmentation of dynamic human action. They showed infants a continuous stream of human object-oriented action featuring three actions (e.g., "look at bottle", "poke finger in bottle", "tilt bottle"). One pair of actions was always “packaged” by a narration overlay (e.g., during "look" and "poke", infants heard an infant-directed utterance. *Wow, do you see what she's doing? She's blixing!*), and one pair was always unpackaged (e.g., during "poke" and "tilt", no narration was provided). On test, infants older than 9.5 months discriminated between sequences that had previously been packaged and those that had not, preferring to look at the novel “unpackaged” sequences [35]. These findings may indicate that acoustic packaging binds units of actions into larger coherent units, which infants then process differently from units that appear without such acoustic cues.

Brand and Tapscott's findings clearly indicate that infants' processing of action can be influenced by the nature of the multimodal input they receive while witnessing human goal-directed activity. One important question arises from these findings, however: To what degree are infants *typically* exposed to the type of acoustic packaging used in Brand and Tapscott's experimental manipulation? That is, is synchronous acoustic packaging an ecologically valid phenomenon that is a regular part of infants' mutimodal input? Gogate and colleagues' work suggests that gesture-speech synchrony is fairly common in mother-child interactions, but their studies addressed this phenomenon within the context of a structured naming task rather than during demonstrations of goal-directed *action*. Thus, it is still an open question as to whether such synchrony is found during action demonstration.

It may seem self-evident that parents would be most likely to discuss actions at the time the actions are being performed. However, there is some evidence suggesting that, at least with older children, actions and the utterances that parents use to describe them are actually not aligned. In a study of verb learning in children in their second year (15-21 months of age), Tomasello and Kruger found that mothers were more likely to provide a given target verb *before* an action was performed. In a subsequent experimental study, the authors also demonstrated that children’s productive verb learning was best when they were provided with a novel verb in advance of an action, exactly the type of timing seen most often in the observational study [36].

On the other hand, there is also some evidence that suggests that acoustic packaging may indeed be characteristic of parents' action demonstrations in younger populations.

Schillingmann, Wrede, and Rohlfing used an automated means of detecting both speech units (via a speech recognizer) and action units (defined as motion taking place between two motion minima, detected via pixel change). They examined footage of parents interacting with their preverbal infants during action demonstration to assess the degree of speech-action alignment. In particular, they examined acoustic packaging, operationalized in their study as periods of speech and action overlap and annotated automatically. Applying this method to both infant-directed and adult-directed action (specifically, a stacking cups task), the authors discovered that infant-directed action featured more acoustic packages and further that infant-directed packages contained fewer motion elements in comparison to adult-directed packages [12]. In another study of naturalistic infant-directed action, Rolf, Hanheide, and Rohlfing similarly assessed signal-level acoustic packaging in multimodal input, defined as temporal correlation between audio and visual signal flows. Consistent with the results of Schillingmann and colleagues, Rolf et al. found more audio-visual correlation in child-directed action demonstrations in comparison to adult-directed demonstrations [13].

IV. Study Overview

The work of Rohlfing and others on multimodal synchrony has provided valuable information regarding the nature of parents' infant-directed modifications to their action and speech integration. The current study adds to these findings by providing descriptive analyses of infant-directed action using a different measure of synchrony. Specifically, whereas Rohlfing and colleagues have reported on speech and action overlap or correlation, the current study provides a similar measure but also assesses the degree to which action onsets and offsets are temporally *aligned* with speech onsets and offsets. That is, rather than defining and detecting synchrony as overlap between speech and action, we assessed the average temporal delays between speech onset and action onset, as well as speech offset and action offset. This provides a measure of synchrony more analogous to that of Gogate and colleagues, who reported on average temporal characteristics of label onsets with respect to gesture onsets.

Information about multimodal temporal alignment as it typically exists in mothers' action demonstrations is important in two respects; first, it can be compared to past studies addressing alignment for other communicative purposes such as labeling (e.g., [34]), and second, it can be used to inform experimental studies of the effects of synchronous acoustic packaging such as that used in Brand and Tapscott [35]. With these goals in mind, for the current study we asked mothers to demonstrate two tasks, stacking rings and nesting cups. We then annotated onsets of actions and speech units to investigate the temporal profiles of these two types of input with respect to each other.

Our study also differs from past studies of acoustic packaging in another important way. Previous studies have approached issues of multimodal synchrony from a robotics perspective and were thus primarily concerned with basic perceptual input (i.e., either speech vs. non-speech and its integration with objective motion change). Clearly, this approach is the most sensible one given the ultimate goal of creating a system from the ground up that can recognize (and perhaps imitate) actions. In contrast, we addressed speech-action alignment from a developmental psychological perspective, and hence made some different starting assumptions about how top-down knowledge might be relevant in the role of multimodal synchrony.

More specifically, past studies (e.g., [12], [13]) assessed audio and visual information on a low level, assuming no prior knowledge about the content of the actions or the semantic differences among utterances. Further, the researchers used these analyses to compare infant- versus adult-directed demonstrations. We, however, examined the patterns of alignment within infant-directed demonstrations, investigating the semantic content of mothers' utterances and making comparisons of temporal speech-action integration based on these semantic classes. We also report on the proportion of utterances overlapping with action demonstrations based on utterance type. That is, we examined how often action-describing utterances happen at the same time as actions in comparison to non-action-describing utterances.

We also used human judgments of action boundaries (i.e., onsets and offsets), rather than relying on automated detection of motion change. This means that we relied on top-down knowledge regarding the contents of mothers' actions, further distinguishing our approach from that of a fully bottom-up model. Given that we are primarily interested in acoustic packaging as it is developmentally relevant, the decision to use adult judgments is appropriate as these represent the endstate of interest. It should be noted, however, that past research has suggested that overall motion change does correlate with human judgments of segment boundaries (e.g., [37]-[39]) and thus we are likely identifying similar units to those of past studies of acoustic packaging.

The decision to focus on semantically-based differences is motivated by the idea that such distinctions may be relevant to the infant learner. Of course, we do not assume that the infant learner can comprehend all or even most of the individual words in parents' utterances. However, given that prior research has demonstrated distinctive prosodic contours corresponding to several basic semantic messages [16], and further that even very young infants can distinguish and respond to these different sound profiles [18], we argue that infants may indeed be capable of making use of action-specific speech and its integration with action. To support this possibility, we conducted a final study examining whether adults can distinguish and identify action-describing and non-action-describing utterances when they were low-pass filtered (thus rendering the individual words incomprehensible). This type of study has been used to demonstrate the distinctive contours of other types of infant-directed speech (e.g., [16], [17]), and thus we add to this body of literature by suggesting that action-describing speech may similarly be distinguishable and recognized as conveying action-related content.

We had three predictions. First, we expected that action-describing utterances would be more synchronous with the actions themselves in comparison to non-action-describing utterances (Analysis 1). This would be reflected by action onsets being temporally closer to action-describing utterance onsets (in comparison to non-action-describing utterance onsets), as well as action offsets being temporally closer to action-describing utterance offsets (in comparison to non-action-describing utterance offsets). Second, we also expected that action-describing utterances would be more likely to overlap with the actions themselves in comparison to non-action-describing utterances (Analysis 2). Finally, we predicted that adults would be able to identify action-describing utterances in comparison to non-action-describing utterances based on prosodic aspects alone, suggesting that the special timing profiles of action-describing speech (if found) might be relevant to the infant learner (Analysis 3).

V. METHOD

A. Participants

We selected a broad age range (6-13 months) so as to enable developmental comparisons. The final sample consisted of 10 younger infants (range 6.0-9.5 months, *M* = 7.7 mo., *SD* = 1.0) and 10 older infants (range 9.6 -13.0 months, *M* = 11.33 mo., *SD* = .75). Gender of infant was roughly equivalent in both groups. All parent-infant dyads contained the infant's mother.

B. Experimental Set-Up and Procedure

Mothers were seated directly across from their infant, who was secured in a high chair with a tray at chest level. A camera was positioned to capture mothers' motions from the side and recorded at a frame rate of 30 FPS (See Figure 1).

We asked mothers to demonstrate a series of three toys to their infant: A ball with suction cups that could grab onto smooth surfaces, a stack of rings, and a series of nesting cups. The ball was used as a warm-up toy; footage from this toy was not analyzed. The ball was always presented first, and the order of the second and third toy (rings or cups) was counterbalanced such that half of the infants saw rings first, and the other half saw cups. We instructed mothers to demonstrate toys to their infants much as they would at home, with the exception that they take time to specifically demonstrate the stacking or nesting function before providing the infant an opportunity to interact with the toys.

C. Coding

From each dyad, we isolated the first demonstration for each toy in which mothers showed the entire toy (i.e., stacked all the rings or nested all the cups) without interruption from the infant. We then coded the footage for action onsets and offsets as well as speech onsets and offsets. Action and speech were coded separately so as not to bias the coders; action was coded with the sound muted, and speech was analyzed using the waveform and sound only, with no visual access to the interaction.

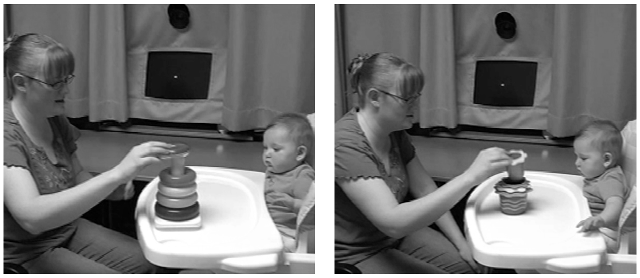

Figure 1: Mother demonstrating ring stacking and cup nesting to infant

A trained coder identified action onsets and offsets frame-by-frame based on the following definitions: An onset was considered to be the moment at which a component (either a ring or cup) was moved down from an apex onto the assembly, and an offset was the moment at which the component was fully placed on or in the main assembly and stopped moving.

Speech was analyzed using Praat, a freely available speech analysis software program [40]. Onsets and offsets of utterances were noted using a combination of visual and aural inspection of the waveform. Pauses of length greater than 200 ms were considered utterance boundaries. Thus, we defined utterances on the basis of pauses rather than semantic content. Semantic judgments were made after the utterances had been located and times of onset and offset recorded. Utterances were typically short, containing two or three words (though longer utterances were also observed); mean duration of utterance was 866 ms (*SD* = 580). Off-task utterances were not transcribed or further coded.

Based solely on inspection of the transcribed lexical content of utterances, a coder identified four semantic utterance classes[1]: *goal setting* (anticipatory descriptions of actions or endstates that had not yet occurred), *completion* (statements about actions or endstates that had occurred), *attention-getting* (utterances designed to recruit the attention of the infant), and *action description* (statements relating to ongoing action). Utterance type was not mutually exclusive; an utterance could receive more than one code. For example, if a mother both recruited her infant's attention by calling the infant's name *and* described an action within the same utterance, this utterance would receive codes for both attention getting and ongoing action description. These utterances were then classified as *mixed.* A second coder assigned these five semantic codes to a randomly-selected 20% of the utterances to ensure reliability. Agreement was high (Cohen's kappa = .87). Examples of each type of statement, as well as the total percentage with which each type of utterance appeared in the entire set of utterances, are provided in Table 1.

[1] We use the term 'semantic class' to denote differentiations based on overall meaning inferred by the experimenters' coding of parental speech; however, past studies of infant-directed speech typically refer to infants' capacity to recognize different *affective* messages (e.g., [16]), and one could also refer to such distinctions as pragmatically-based, given that their interpretation is context-bound and not equivalent to the formal lexical meanings of the utterances. For consistency, however, we continue to use the term 'semantic.'

TABLE 1
EXAMPLES AND PERCENTAGES OF UTTERANCE TYPES

Utterance Type	Example	%
Action Description	*Put the blue one on*	57.3
Goal Setting *	*Let's put 'em back on*	7.3
Completion *	*Yaaay, we did it!*	1.9
Attention Getting *	*(Child's name), look!*	21.4
Mixed*	*(Child's name), look it goes in!*	12.2

* = Non-action Description

VI. RESULTS

For every utterance onset, we located the nearest action onset and calculated the (absolute) temporal distance between the two (see Figure 2) to provide a measure of *onset synchrony*; similarly, for every utterance offset, we located the nearest action offset and again calculated temporal distance for a measure of *offset synchrony*.[2] For the sake of simplicity and because some semantic classes of utterances were very infrequent (e.g., completion utterances), we collapsed goal setting, completion, attention recruitment, and mixed utterances into a *non-action description* class and compared their timing with *action descriptions.*

We conducted a mixed between-within ANOVA with age group (younger vs. older) as between-subjects and semantic class of utterance (action description vs. non-action description) as within-subjects to compare onset synchrony (as defined by average absolute temporal distance between utterance onsets and action onsets). We found a main effect for semantic class; as predicted, *action description* onsets were more synchronous (i.e., temporally closer) to the onsets of action units themselves (*M* = .83 sec, *SD* = .33) in comparison to *non-action description* onsets (*M* = 1.49 sec, *SD* = .64), *F* (1, 18) = 18.29, *p* < .001. There was no effect for age, nor was there an age x semantic class interaction (*p*'s > .1). When examining offset synchrony, we again found the predicted main effect for semantic class; offsets of *action descriptions* were temporally closer with offsets of the actions themselves (*M* = .81 sec, *SD*= .35) in comparison to *non-action description* offsets (*M* = 1.41 sec*, SD* = .64), again with no interaction (see Figure 3).

We also conducted a second analysis to provide more descriptive information about mothers' acoustic packaging, calculating the proportion of action and non-action descriptions that overlapped with action units. Overlap was defined as any period of time in which action and speech were co-occurring; thus an utterance was considered overlapping if any portion of it co-occurred with an action. This analysis provides information similar to that of previous studies of acoustic packaging (e.g., Schillingmann et al. [12]), which demonstrated that more overlap occurred during infant-directed vs. adult-directed action. Here, however, the comparison was between overlaps of action-describing utterances vs. non-action-describing utterances, with the prediction of more overlaps during the the former.

We ran a mixed between-within ANOVA with age group (younger vs. older) as between-subjects and semantic class of utterance (action description vs. non-action description) as within-subjects to compare proportion of utterances that overlapped with action units. As predicted, there was a main effect for semantic class, with average proportion of action-describing utterances overlapping with actions significantly higher (*M* = .55, *SD* = .27) than proportion of non-action-describing utterances overlapping with actions (*M* = .28, *SD* = .19). There was also a significant effect for age; mothers of older infants had an overall higher proportion of overlapped utterances (*M* = .47, *SD* = .25) in comparison to mothers of younger infants (*M* = .36, *SD* = .28). The utterance type x age interaction was not significant (*p* > .1) (see Figure 4).

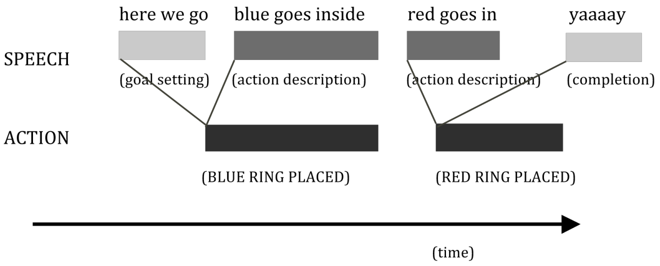

Figure 2: Schematic representation of the speech and action stream for calculating onset synchrony. Onset synchrony for these units would be assessed by comparing the temporal distance between *action description* onsets and action onsets vs. *non-action description* onsets (here a goal setting and a completion) and action onsets. A similar calculation was performed for offsets. In this example, the onsets of the non-action descriptions ('here we go' and 'yaaay') are relatively far in time from the onsets of the actions of placing the blue and red rings respectively; the onsets of the action descriptions ('blue goes inside' and 'red goes in' are closer in time to the onsets of blue and red ring placement.

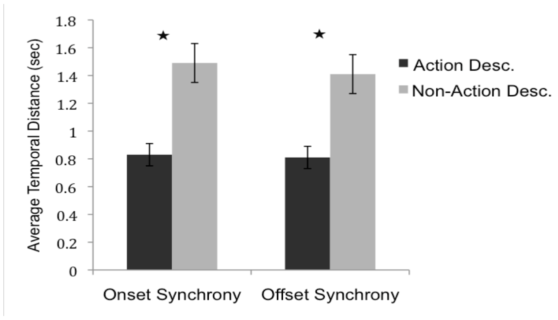

Figure 3: Average temporal distance by utterance type (action descriptions vs. non-action descriptions). Action description utterances were aligned more closely with action units both in their onsets and their offsets, demonstrating more synchrony for action-describing utterances. Standard error bars represent ± 1 SE of the mean.

[2] With respect to sequencing of utterances and actions, there was a slight tendency for utterance onsets to precede action onsets; the majority (59%) of action description utterances started before action started. There was also a slight tendency for utterance offsets to come after action offsets; 56% of action description utterances ended after action ended. However, these differences were not significant by binomial tests, *p*s > .05

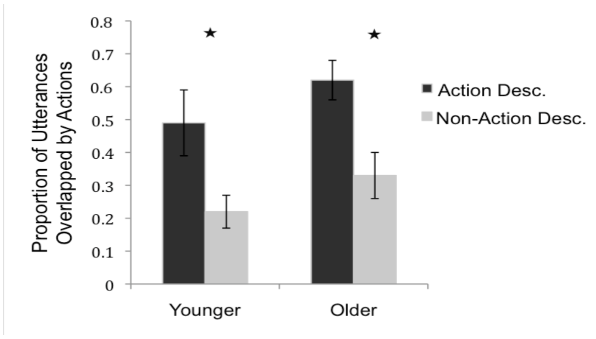

Figure 4: Average proportion of action and non-action descriptions that overlapped with action units. Action descriptions were proportionately more overlapped with actions in comparison to non-action descriptions. Standard error bars represent ± 1 SE of the mean.

Thus far, we have focused on analyses targeted at the temporal profiles of action-describing vs. non-action-describing speech. A final, third analysis examined whether the prosodic features of action-describing utterances were discriminably different from non-action-describing utterances.

We extracted five action-description utterances and five utterances from non-action-description utterances, using utterances from nine of the mothers. Utterances were roughly matched on duration (and did not differ significantly according to a *t* test, $p > .05$). Each utterance was then low-pass filtered (to 450 Hz) using Audacity software [41]. Adult subjects (2 males, 8 females) listened to all ten clips in one of two random orders and judged whether they were action descriptions or other types of utterances. Subjects correctly identified utterance types at a level significantly greater than chance, $M = 67\%$, $SD = 16$, one-sample $t\ (9) = 3.29$, $p = .009$. When asked to identify any words from the clips, subjects identified only 6% of the possible tokens, suggesting that the words were indeed largely unintelligible and that the filtering was successful in masking articulatory features.

VII. Discussion

Our study demonstrated that mothers aligned their speech and action during action demonstration with their infants. Specifically, more temporal synchrony was observed between actions and action-describing utterances in comparison to other types of utterances. That is, both onsets and offsets of action-describing speech tended to be temporally closer to action unit onsets and offsets, respectively. It is also the case that there were proportionately more instances of temporally co-occurring (overlapping) utterances and action when those utterances were action-describing. Taken together, these findings suggest that mothers structure their interactions to provide multimodal information directly relevant to the actions being demonstrated. The findings also lend support to the idea that acoustic packaging of action is a commonplace source of multimodal information.

In general, we did not reveal developmental effects when examining whether acoustic packaging differs for different age groups. However, we did find that mothers *overall* tended to overlap both action-describing and non-action-describing speech with actions more with their older infants. We do not take a strong position on why this effect was obtained; however, we speculate that it may be due to mothers of older and thus more exploratory children trying to increase their infant's attention to their own demonstrations in an attempt to distract them from exploring and trying the toys on their own.

The utility of the semantic-specific alignment that we found would of course only be relevant to the infant learner if the infant could actually discriminate the different types of utterances as well as respond to the different semantic (or affective/pragmatic) messages being conveyed. In a first step to argue for such an ability, our final analysis demonstrated that the prosodic contours of action-describing and non-action-describing utterances were sufficiently different to allow for discrimination and categorization by adults. Namely, adults were able to classify low-pass filtered utterances as action-describing or non-action-describing at above-chance levels. Our findings parallel those of other similar studies, including Fernald's [16] as well as Bryant and Barrett's [17] demonstrations of adults' categorization of other types of infant-directed speech. Studies such as these are an important initial demonstration of how various global semantic/pragmatic functions can be expressed in infant-directed speech. Future research can address whether infants themselves discriminate between action-describing vs. non-action-describing utterances, as well as provide more objective information about the prosodic contours specific to action description (e.g., pitch and pitch excursions).

By taking an approach in which semantic content was used as a basis for synchrony calculations, our method most closely resembles that of Gogate and colleagues (e.g., [34]), who demonstrated that mothers' labeling words tend to be more synchronous with movement of the relevant object in comparison to other types of words. Our own results argue for a similar finding within a different type of interaction, specifically that maternal action descriptions tend to be more synchronous with actions than other types of utterances. In this way, our findings complement Gogate and colleagues, suggesting that infant-directed communication may feature structured speech and action for a range of purposes (e.g., labeling, action demonstration).

Still unanswered in our analyses is the extent to which our findings are specific to infant-directed communication. It is possible that the alignment of action-describing speech and action is simply a byproduct of normal speech; as noted above, it is an entirely intuitive finding that a speaker's utterances should concern ongoing events, so it is sensible that mothers would refer to actions as they performed them (though see [36] for evidence to the contrary). Note that this "byproduct" account, if true, would *not* imply that alignment is useless for the infant learner; it would simply imply that the alignment is not an infant-specific modification to maternal communication. It would, however, still be useful in the future to examine footage of mothers demonstrating similar actions in dyadic interactions with adults to assess any changes in either synchrony or overlap, as well as changes to action-describing utterance prosody. These lines of investigation create inviting opportunities for future research.

The analyses we present in the current paper are novel in at least one important way. Although previous research has provided information on how parents adapt their speech and action to infants relative to adults ([12] - [16], [23], [24]), to date there has been little investigation *within* infant-directed action demonstrations focusing on the frequency and nature of various semantic classes of utterances. In this regard, our study represents the first assessment of acoustic packaging with respect to how different *types* of utterances are used to provide infants with multimodal speech and action information. Given that infants may be capable of discriminating utterances on the basis of prosodic characteristics, we suggest that mothers' specific alignment of action-describing utterances may provide infants an integrated speech-action signal capable of facilitating infants' attention to relevant action units. This acoustic packaging, then, may well be part of the solution to how infants come to process and understand events.

REFERENCES

[1] D. A. Baldwin and J. A. Baird, "Discerning intentions in dynamic human action," *Trends Cogn. Sci.*, vol. 5, pp. 171-178, Apr. 2001.

[2] D. Newtson and G. Engquist, "The perceptual organization of ongoing behavior," *J Exp. Soc. Psychol.,* vol. 12, pp. 436-450, Sept. 1976.

[3] N. K. Speer, K. M. Swallow, and J. M. Zacks, "Activation of human motion processing areas during event perception," *Cogn. Affect. Behav. Neurosci.,* vol. 3, pp. 335-345, Dec. 2003.

[4] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, et al., "Human brain activity time-locked to perceptual event boundaries," *Nat. Neurosci.*, vol. 4, pp. 651–655, June 2001.

[5] J. M. Zacks and K. M. Swallow, "Event segmentation," *Curr. Direct. Psychol. Sci.*, vol. 16, pp. 80-84, Apr. 2007.

[6] B. M. Hard, B. Tversky, and D. Lang, "Making sense of abstract events: Building event schemas," *Mem. Cognit.*, vol. 34, pp. 1221-1235, June 2006.

[7] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures.* Hillsdale, NJ: Erlbaum, 1977.

[8] D. A. Baldwin, J. Baird, M. M. Saylor, and M. A. Clark, "Infants parse dynamic action," *Chi. Dev.*, vol. 72, pp. 708–718, May/June 2001.

[9] M. M. Saylor, D. A. Baldwin, J. A. Baird, and J. LaBounty, "Infants' on-line segmentation of dynamic human action," *J. Cognit. Dev.*, vol. 8, pp. 113-128, Feb. 2007.

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

[11] P. Zukow-Goldring. "Socio-perceptual bases for the emergence of language: An alternative to innatist approaches." *Dev. Psyo. Biol.,* vol. 23, pp. 705-726.

[12] L. Schillingmann, B. Wrede, and K. J. Rohlfing, "A computational model of acoustic packaging," *IEEE Trans Autonomous Mental Dev.* vol. 1, pp. 226-236, Dec. 2009.

[13] M. Rolf, M. Hanheide, and K. J. Rohlfing, "Attention via synchrony: Making use of multimodal cues in social learning," *IEEE Trans Autonomous Mental Dev*, vol. 1, pp. 55-67, May 2009.

[14] C. E. Snow and C. A. Ferguson, Eds., *Talking to Children: Language Input and Acquisition.* Cambridge, England: Cambridge University Press, 1977.

[15] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nat. Rev. Neurosci.,* vol. 5, pp. 831-843, Nov. 2004.

[16] A. Fernald, "Intonation and communicative intent in mother's speech to infants: Is the melody the message?" *Chi. Dev.,* vol. 60, pp. 1497-1510, Dec. 1989.

[17] G. A. Bryant and H. C. Barrett. "Recognizing intentions in infant-directed speech," *Psychol Sci,* vol. 18, pp. 746-751, Aug. 2007.

[18] A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages," *Chi. Dev.*, vol. 63, pp. 657-674, June 1993.

[19] E. D. Thiessen, E. Hill, and J. R. Saffran, "Infant-directed speech facilitates word segmentation," *Infancy,* vol. 7, pp. 53-71, Feb. 2005.

[20] H. Bortfeld, J. Morgan, R. Golinkoff, and K. Rathbun, K. "*Mommy* and me: Familiar names help launch babies into speech stream segmentation," *Psychol. Sci.,* vol. 16, pp. 298-304, Apr. 2005.

[21] P. W. Jusczyk, K. Hirsh-Pasek, D. G. Kemler Nelson, L. Kennedy, A. Woodward, and J. Piwoz, "Perception of acoustic correlates of major phrasal units by young infants," *Cog. Psychol.*, vol. 24, pp. 252-293, Apr. 1992.

[22] A. Fernald, "Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *The Adapted Mind,* J. H. Barkow, L. Cosmides, and J. Tooby, Eds. New York, NY: Oxford University Press, 1992, pp. 391-428.

[23] R. J. Brand, D. A. Baldwin, and L. A. Ashburn, "Evidence for 'motionese': Modifications in mothers' infant-directed action," *Develop. Sci.*, vol. 5, pp. 72–83, Mar. 2002.

[24] R. J. Brand, W. L. Shallcross, M. G. Sabatos, and K. P. Massie, "Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant- versus adult-directed action," *Infancy,* vol. 11, pp. 203-214, May 2007.

[25] Y. Nagai and K. J. Rohlfing, "Can motionese tell infants and robots 'What to imitate'?" in *Proc. 4th Int. Symp. on Imitation in Animals and Artifacts,* 2007, pp. 299-306.

[26] J. Schmidt, J. Fritsch, and B. Kwolek, "Kernel particle filter for real-time 3d body tracking in monocular color images," in *Proc. Automatic Face and Gesture Recognition*, 2006, pp. 567–572.

[27] 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, pp. 1183-1199, 2006.

[28] P. K. Kuhl and A. N. Meltzoff, "Speech as an intermodal object of perception," in *Perceptual Development in Infancy: The Minnesota Symposia on Child Phonology,* vol. 20, A. Yonas, Ed. Hillsdale, NJ: Erlbaum, 1988, p. pp. 235-266.

[29] K. MacKain, M. Studdert-Kennedy, S. Speker, and D. Stern, "Infant intermodal speech perception is a left-hemisphere function," *Science,* vol. 219, pp. 1347-1349, Mar. 1983.

[30] L. Bahrick, R. Lickliter, and R. Flom, "Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy," *Curr. Direct. Psychol. Sci.*, vol. 13, pp. 99-102, June 2004.

[31] L. Bahrick and R. Lickliter, "Intersensory redundancy guides attentional selectivity and perceptual learning in infancy," *Dev. Psychol.,* vol. 36, pp. 190-210, Mar. 2000.

[32] L. J. Gogate and L. E. Bahrick, "Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations," *Infancy*, vol. 2, pp. 219–231, May 2001.

[33] D. A. Baldwin, "Infants' ability to consult the speaker for clues to word reference," *Journal of Child Language,* vol. 20, pp. 395-418, June 1993.

[34] L. Gogate, L. Bahrick, and J. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child Develop.*, vol. 71, pp. 878–894, Jul./Aug. 2000.

[35] R. J. Brand and S. Tapscott, "Acoustic packaging of action sequences by infants," *Infancy*, vol. 11, pp. 321–332, May 2007.

[36] M. Tomasello and A. Kruger, "Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts," *J. Child Lang,* vol. 19, pp. 311 – 333, June 1992.

[37] B. Hard, "Reading the language of action: Hierarchical encoding of observed behavior," Doctoral dissertation, Stanford University. 2006.

[38] B. Hard and G. Recchia, "Reading the language of action," in N.A. Taatgen and H. van Rijn, Eds., *Proc. 28th Annual Conf. of the Cognitive Science Society*, 2006, pp. 1433-1439.

[39] M. Meyer, P. DeCamp, B. Hard, D. A. Baldwin, and D. Roy, "Assessing behavioral and computational approaches to naturalistic action segmentation," in S. Ohlsson & R. Catrambone, Eds., *Proc. 32nd Annual Conf. of the Cognitive Science Society*, 2010, pp. 2710-2715.

[40] P. Boersma and D. Weenik, "Praat: Doing phonetics by computer (Version 5.1.05) [Computer program]," Retrieved May 2009, from http://www.praat.org

[41] D. Mazzoni and R. Dannenburg, "Audacity (Version 1.2.6.a)," Retrieved July 2010 from http://audacity.sourceforge.net/

Meredith Meyer received the Master's and Ph. D. degrees in developmental psychology from the University of Oregon, OR (United States), in 2009. In 2010 she began a position as a postdoctoral researcher at the University of Michigan. She researches issues related to child and adult action processing, statistical learning of action, gesture comprehension, and language development, with a current focus on how language and gesture impact language comprehension and concept development.

Bridgette Hard received a Ph.D. in Cognitive Psychology from Stanford University, CA (United States), in 2006. She then received a National Research Service Award from the National Institute of Health to pursue postdoctoral research in developmental psycholgy at the University of Oregon from 2006 to 2009. In 2009, she became the coordinator of the Psychology One Program at Stanford University. Her research focuses on how adults and children segment and understand ongoing action.

Rebecca J. Brand is an Associate Professor of Psychology at Villanova University, PA (United States). She received her Bachelor's degree in Cognitive Science from Vassar College and her Master's and Ph.D. in Psychology from the University of Oregon. Her research focuses on infants' social-cognitive skills and the role of parental input in the development of these skills. Current work investigates infant-directed action ("motionese") as well as speech-action alignment ("acoustic packaging").

Dare A. Baldwin completed undergraduate work in Psychology at U.C. Berkeley, received a master’s at U.C. Santa Cruz, and a Ph.D. in Psychology, with a special designation in Cognitive Science, at Stanford University, CA (United States).

She is now Professor of Developmental Psychology at the University of Oregon. Her research focuses on enhancing the human potential for knowledge acquisition and in understanding how human infants rapidly and effectively acquire and organize world knowledge across many domains.

A past fellow at the Center for Advanced Study in the Behavioral Sciences, Dr. Baldwin is currently a fellow of the Association for Psychological Science, and is on the executive boards of the Society for Language Development and the Cognitive Development Society.

Molly McGarvey is a graduate student at Villanova University, PA (United States). She received her Bachelor’s degree in English with a minor in Psychology from Ohio University and will graduate with her Master’s in Psychology in May 2010. She is studying adult and child reactions to acoustic packaging, and she plans to go on to a doctoral program in clinical or developmental psychology.