Infants temporally coordinate gesture-speech combinations before they produce their first words

Núria Esteve-Gibert[1*], Pilar Prieto[2, 1]

[1]*Dpt. of Translation and Language Sciences, Universitat Pompeu Fabra, Spain*

E-mail: nuria.esteve@upf.edu

Postal address: Roc Boronat, 138, 08018 Barcelona (Spain)

[2] *Institució Catalana de Recerca i Estudis Avançats, ICREA*

* Corresponding author
Contact details of the corresponding author:
Tel.*:* 0034-935422409
Fax*:* 0034-935421617
Postal address: Roc Boronat, 138, 08018 Barcelona, Spain
E-mail nuria.esteve@upf.edu

ABSTRACT

This study explores the patterns of gesture and speech combinations from the babbling period to the one-word stage and the temporal alignment between the two modalities. The communicative acts of four Catalan children at 0;11, 1;1, 1;3, 1;5, and 1;7 were gesturally and acoustically analyzed. Results from the analysis of a total of 4,507 communicative acts extracted from approximately 24 hours of at-home recordings showed that (1) from the early single-word period onwards gesture starts being produced mainly in combination with speech rather than as a gesture-only act; (2) in these early gesture-speech combinations most of the gestures are deictic gestures (pointing and reaching gestures) with a declarative communicative purpose; and (3) there is evidence of temporal coordination between gesture and speech already at the babbling stage because gestures start before the vocalizations associated with them, the stroke onset coincides with the onset of the prominent syllable in speech, and the gesture apex is produced before the end of the accented syllable. These results suggest that during the transition between the babbling stage and single-word period infants start combining deictic gestures and speech and, when combined, the two modalities are temporally coordinated.

Keywords: early gestures; early acquisition of multimodality; early gesture-speech temporal coordination

1. INTRODUCTION

There is a broad consensus in the literature on the tight relationship and mutual influence between gesture and speech. Many researchers have stated that gesture and speech form an integrated system in communication (De Ruiter, 2000; Kendon, 1980; Kita, 2000; McNeill, 1992). Important features that back up the speech-gesture integration analysis in adults are that most of the gestures are produced together with speech, and that the two modalities are (a) semantically and pragmatically coherent, and (b) temporally synchronized, i.e., the most prominent part of the gesture is temporally integrated with speech (McNeill, 1992).

Studies on the temporal coordination of gesture and speech provide strong evidence for the claim that gesture and speech form an integrated system in adults. It has been shown that the most prominent part of the gesture typically co-occurs with the most prominent part of the speech (Kendon, 1980). But different anchoring regions in speech have been proposed to serve as coordination sites for gesture prominence locations: speech onset (Bergmann, Aksu & Kopp, 2011; Butterworth & Beattie, 1978; Ferré, 2010; Levelt, Richardson & La Heij, 1985; Roustan & Dohen, 2010), prosodically prominent syllables (Krahmer & Swerts, 2007; Loehr, 2007; Leonard & Cummins, 2010), or prosodically prominent syllables with intonation peaks (De Ruiter, 1998; Esteve-Gibert & Prieto, in press; Nobe, 1996). Taking together these findings, there is general agreement in the literature that (a) prominences in gesture and speech are temporally coordinated, (b) the prominence in gesture is represented by the gesture stroke (in the case of a deictic gesture, the interval of time during which the arm is extended) or the

gesture apex (the specific point within the stroke interval at which the finger is maximally extended), and (c) the prominence in speech is represented by the prosodically prominent syllable. In our study, these measures will be taken into account in order to investigate the development of the temporal coordination of gesture with speech in the early stages in language development.

But are gestures coordinated with speech in young infants to form an integrated system the way that they are in adults? Iverson and Thelen (1999) and Masataka (2003) state that speech and gesture coordinations have their developmental origins in early hand-mouth linkages. Based on the dynamic systems theory of motor control, they propose that systems activating mouth and arms can influence and entrain one another, and these entrainments are dynamic and flexible such that activation in one system can affect the other in the form of a looser or tighter temporal synchrony. However, a given behaviour must be strong and stable (with low threshold and high activation) to pull in and entrain the activity of the complementary system. Iverson and Thelen (1999) propose four developmental stages of dynamic progression, namely, (1) in newborns, an early oral-manual system in which instances of hand-mouth contact and co-occurrences of hand movements with vocalizations are frequent; (2) from 6 to 8 months, rhythmical movements with the hands and mouth showing an increasing control over the manual and oral articulators, and possibly indicating the transition into the speech-gesture system; (3) from 9 to 14 months, a more articulated control over the two modalities, which are then more directed to communication, with the gesture modality predominating but with entrainment also occurring between the two, and showing a tight relation between early gesture production and a later language development, and

(4) from 16 to 18 months, a tighter control over both modalities, leading to the emergence of synchronous gesture and speech combinations.

In order to investigate deeply the temporal overlap between the occurrence of vocalizations and rhythmic activities of the limbs in infants, Ejiri and Masataka (2001) investigated the vocal and motor behaviour of 4 Japanese infants from 0;6 to 0;11. In a first study, they examined the temporal overlap between vocalizations and rhythmic activities during the babbling stage. The authors found that vocalizations very frequently co-occurred with rhythmic actions, and interestingly that these coordinated behaviours increased immediately before and during the month in which canonical babbling was initiated. In a second study, they compared vocalizations co-occurring with rhythmic actions to vocalizations not co-occurring with rhythmic actions, and they found that syllable and formant frequency durations were shorter in vocalizations co-occurring with rhythmic actions than in non-co-occurring ones.

Similarly, Iverson and Fagan (2004) described early infants' production of vocal-rhythmic movement coordination by testing 47 infants between the ages of 0;6 and 0;9. Results showed that at 0;7 vocal-motor coordination was a stable component of infants' behavioural repertoires, and that these early coordinations were a developmental precursor to the gesture-speech system. The authors based this statement on three observations: (1) infants at all ages coordinated vocalizations with single-arm rhythmic movements more often than with both-arm movements; (2) at all ages the proportion of coordinated right-arm movements was higher than that of left-arm movements, paralleling adult-like behaviours; and (3) most of the coordinations followed the

temporal patterns of organizing gesture-speech productions, since motor activities were synchronous with or slightly anticipated vocalization onsets.

The abovementioned studies focusing on rhythmic movements revealed that vocal and motor rhythmic movements are precursors of the coordination between the gesture and speech modalities. However, very few studies have investigated the patterns of that early coordination itself, i.e., the early synchronization between communicative gestures and vocalizations. To our knowledge, only Butcher and Goldin-Meadow (2000) have performed such a study. The authors analyzed six children longitudinally in spontaneous play situations during the transition from one- to two-word speech in order to investigate whether (1) at that age children produce gestures with or without speech, (2) children temporally coordinate gesture and speech, and (3) children semantically integrate the two modalities. First, they found that the production of utterances containing gesture remained stable across the ages analyzed, but with a difference between age groups: at the beginning of the single-word period gestures were generally not accompanied by speech, and at the end of the single-word period children mainly combined them with speech. Second, they found that it was not until the beginning of the two-word period that children produced gesture-speech combinations in which the speech co-occurred with the most prominent part of the gesture (defined by them as the stroke or peak of the gesture, i.e., the farthest extension before the hand began to retract). Finally, the study showed that the proportion of gestures produced in combination with meaningful speech (as opposed to *meaningless* speech) increased across the ages analyzed. In conclusion, Butcher and Goldin-Meadow (2000) suggested that it is not until the beginning of the two-word period that children integrate gesture and speech as a single system to communicate intentionally.

The present paper aims at contributing to the research on early coordination between communicative gesture and speech by describing the emergence of gesture-speech combinations and their temporal coordination. Following up on Butcher and Goldin-Meadow (2000), we incorporate two innovative aspects in our study. The first innovative aspect is an analysis of the emergence of communicative gesture-speech combinations starting from the babbling period. The babbling period emerges in the middle of the first year of life and it is a crucial stage in language development because it provides the raw material for the production of early words (Oller, Wieman, Doyle & Ross, 1976; Vihman, Macken, Miller, Simmons & Miller, 1985). In the frame of the dynamic systems theory (Iverson & Thelen, 1999), Vihman, DePaolis and Keren-Portnoy (2009) propose an 'articulatory filter': the first syllables that infants produce when babbling help the bootstrapping of the speech stream, and consequently the development of the phonological systematicity. Thus, during the second half of the first year, infants start practicing very simple, accessible, and accurate syllables. Once these syllables are well practiced, the infants' attention is unconsciously captured by sound patterns in the speech stream that match good enough his or her own babbled productions. Consequently, the child can detect if a sound pattern occurs repeatedly in a given situation and, when experiencing a similar situation, the child will be primed to produce those syllables. According to the authors, this fact can strengthen the memory trace and support the memory for the mapping between form and meaning. Also, the babbling period coincides with the period when communicative gestures start being produced (Bates, Camaioni & Volterra, 1975; Tomasello, Carpenter & Liszkowski, 2007). The second innovative aspect in our study is a more fine-grained temporal coordination analysis that incorporates recent findings on the way gesture and speech

temporally coordinate in adult speech and that takes into account the importance of prosodic prominence in gesture-speech alignment patterns. This will allow us to assess the degree of temporal coordination in more detail.

Thus, the goal of this study is two-fold. First, it aims to describe when and how children combine communicative gestures with speech in the babbling and single-word periods. In order to fulfil this aim, the study will analyze the intentional gesture-speech combinations produced by 4 children between the ages of 0;11, and 1;7, the ages in which children start producing most of their communicative gestures in combination with speech, and then go on to examine the gesture types and motives that appear most frequently in these early gesture-speech combinations. Second, it aims to investigate precisely the early temporal synchronization of gesture with speech. To this end, the study will analyze a variety of measures that have been found useful in recent studies involving adults, as follows: the temporal distance between gesture onset and speech onset; the temporal distance between stroke onset and speech onset; the temporal distance between stroke onset and the beginning of the accented syllable; and the temporal distance between the gesture apex and the end of the accented syllable. We hypothesize that we will find evidence of temporal coordination in early gesture-speech combinations as they emerge in the transition between the babbling and single-word periods.

2. METHOD

2.1 Participants

The participants of the longitudinal study are four Catalan-learning infants, two male (who will be called Bi and Ma) and two female (who will be called An and On). The children are all from middle-class homes in four small towns located within the same region of Catalonia, Alt Penedès, located 50 km to the south of Barcelona. Although varying degrees of bilingualism between Catalan and Spanish exist throughout Catalonia, according to the official statistics website of Catalonia (www.idescat.cat, Linguistic census from 2008) linguistic census, in that region Catalan is spoken regularly by about 81% of the population. All parents of the four participants spoke exclusively in Catalan with their child and to each other. Parents were asked about their linguistic habits through a language questionnaire, and they showed a mean 85% of use of Catalan in their dealings with other family members, friends, and work colleagues.

2.2 Procedure

The children participated in free play activities as they interacted with their caregiver. Caregivers were told to interact naturally with the children, playing as they would in their everyday lives. No other instructions on how to play or interact were given to them. Sessions were videotaped in the subjects' respective homes, typically in the living-room. The experimenter hand-held the camera while recording child and caregiver, and if the child-caregiver dyad moved to another room, the experimenter followed them with the camera.

Recording sessions took place from when children were 0;11 until they were 1;7, either weekly or biweekly, and lasted between 30 and 45 minutes, depending on the attention

span of the children. All recordings have been made public through the Esteve-Prieto Catalan acquisition corpus, which includes recordings of these four children from the age of 0;7 until they were 3;0 (Esteve-Gibert & Prieto, 2012). Recordings were made using a SONY camera, model DCR-DVD202E PAL. No additional microphones other than the one in the camera was attached to the children's clothes or installed in the room. This had the advantage of obtaining more naturalistic data, because infants could move freely around the house and they did not play with a strange object attached to their clothes. However, it also had the disadvantage that the data did not have a perfect acoustic quality. In order to palliate this effect, the author of the recordings tried to be as close as possible to the children without interfering with their activities.

2.3 Data coding

The present study analyzes children's gesture and speech productions at ages 0;11, 1;1, 1;3, 1;5, and 1;7. Children were recorded weekly at 0;11 and biweekly from 1;1 to 1;7. A total of 39 sessions thus yielded a total of approximately 24 hours of video stream. This age range was intended to include the infants' babbling period because it is at this stage that infants produce their first communicative gestures, mostly in the form of pointing and reaching gestures. The age span included in our study therefore constitutes an earlier span than that analyzed in Butcher and Goldin-Meadow (2000), who started analyzing children as soon as they were at the single-word period and finished their analysis when children produced two-word combinations.[1]

[1] Due to individual differences, the infants' ages in Butcher & Goldin-Meadow (2000) varied significantly: one child was analyzed from 1;0 to 2;1, one from 1;1 to 1;7, two children from 1;3 to 1;9, one from 1;3 to 2;1, and another one from 1;9 to 2;3.

Following De Boysson-Bardies and Vihman (1991), the onset of word production was established as the first session in which the child used four or more words (the 4-word point), whereas the first session in which approximately 25 or more words were used spontaneously was identified as the starting point for the single-word period. All four children were at the babbling stage at 0;11 and 1;1 because they were still not producing 4 words during these recording sessions, and all four children were already at the single-word period at 1;5, because at that point they produced 25 or more words during a recording session. Individual differences were found at age 1;3: at this age three infants produced around 4 words during one recording session, and the other infant produced around 20 words per session. Age 1;3 was defined as the onset of word production given that all the children were producing 4 words or more at this point. Table 1 summarizes the number of recorded sessions that were included in the study, classified by the children's age, duration of the sessions, and number of words produced during them. As this table shows, we analyzed 39 recording sessions of about 30 minutes each, which means a total of 24 hours of video recordings.

Participant	*Age*	*Duration*	*Number of words per session*
An	0;11.03	0:33:00	0
	0;11.08	0:36:34	1
	0;11.15	0:36:35	0
	1;1.10	0:37:21	3
	1;1.24	0:41:48	2
	1;3.07	0:29:10	15
	1;3.28	0:34:49	21
	1;5.07	0:25:29	23
	1;5.28	0:33:54	26
	1;7.05	0:35:42	50
	1;7.16	0:34:21	56
Bi	0;11.12	0:36:20	0
	0;11.18	0:34:21	0
	0;11.25	0:26:09	0
	1;1.07	0:34:59	1
	1;1.20	0:34:05	0
	1;3.15	0:35:57	6

	1;3.29	0:35:31	9
	1;5.03	0:37:55	17
	1;5.17	0:37:58	22
	1;7.26	0:37:12	34
Ma	0;11.05	0:34:43	0
	0;11.12	0:39:44	1
	0;11.19	0:35:20	0
	0;11.25	0:33:23	1
	1;1.14	0:31:17	2
	1;1.27	0:33:36	4
	1;3.08	0:35:48	5
	1;3.22	0:32:56	7
	1;5.23	0:34:51	26
	1;7.05	0:36:29	31
On	0;11.14	0:26:25	1
	0;11.23	0:37:12	1
	1;1.06	0:37:50	2
	1;1.28	0:36:15	4
	1;3.08	0:23:28	6
	1;3.21	0:36:43	9
	1;5.15	0:37:09	22
	1;7.14	1:10:54	27
TOTAL	39 sessions	23:16:00	

Table 1. Recorded sessions included in the study, classified by children's age, duration of the session and number of words produced per session.

All communicative acts (visual and/or vocal) produced by the infants were identified and located in the recordings by the first author using the ELAN annotation tool (Lausberg & Sloetjes, 2007). An act was considered to be communicative if (a) the coder perceived or judged the infant's act as based on awareness and deliberate execution (Feldman & Reznick, 1996), if (b) infants produced it in a joint attention frame (either the child directed the gaze to the caregiver before or after the gesture, or the caregiver was attending to what the child was doing), or if (c) the parental reactions before or after the acts suggested so. The adults' perception of the infants' acts as being intentional has been widely used in previous studies as a measure for investigating the children's development of language and cognitive capacities (Butcher & Goldin-

Meadow, 2000; Feldman & Reznick, 1996; Papaeliou & Trevarthen, 2006; Rochat, 2007).

Following So, Demir and Goldin-Meadow (2010), Iverson and Goldin-Meadow (2005), and Özçalişkan and Goldin-Meadow (2005), we excluded from the database all hand movements that involved direct manipulation of an object or were part of a ritualized game. The approximately 24 hours of recordings were thus segmented into 4,507 communicative acts, and then further classified as being 'speech-only' (3,110 instances), 'gesture-only' (668 instances), or a 'gesture-speech combination' (729 instances).

To test the reliability of locating communicative acts and deciding whether they were speech-only, gesture-only, or a gesture-speech combination, two inter-transcriber reliability tests were conducted with a subset of 10% of the data (450 cases) by two independent coders. We made sure that all children and ages were uniformly represented. The overall agreement for the location of communicative acts was 83% and the free marginal kappa statistic obtained was 0.67, indicating that there was substantial agreement between coders regarding the identification and location of communicative acts. The overall agreement for the classification of communicative acts into one of the three categories (namely, speech-only, gesture-only, or gesture-speech combination) was 87% and the kappa statistic was of 0.81, indicating that there was almost perfect agreement between coders.

2.3.1 Speech coding

All infants' communicative acts containing speech were further annotated. First, they were annotated as containing a *vocalization*, if the speech sound conveyed communicative purpose but did not resemble any Catalan word, or a *word*, if the speech sound was clearly a Catalan word or was used consistently. This coding was used to assess the infants' lexical development (see footnote 2). Second, all speech involving simultaneous acts was annotated to determine (a) the limits of the vocalization or word, i.e., its starting and end points (tier 6 in figure 4), and (b) the limits of prosodic prominence, i.e., starting and end points of the accented syllable (tier 7 in figure 4). If the accented syllable of the vocalization or word was not clearly identified, it was coded as an extra category called *fuzzy accented syllable* and excluded from the statistical analyses (N = 65).

Figure 1 shows an example of the acoustic labelling in Praat (Boersma & Weenink, 2012) that was later imported into ELAN and figure 2 summarizes the speech coding conducted.

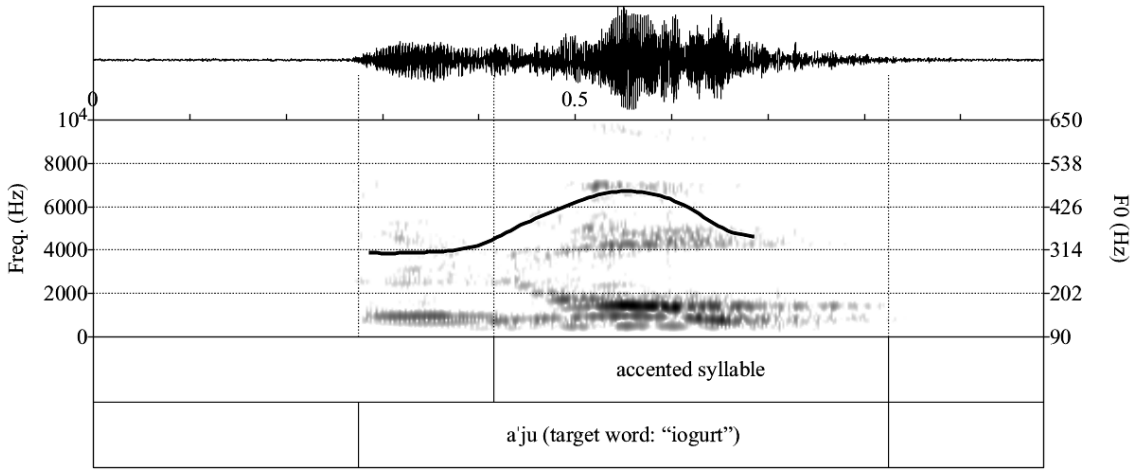

Figure 1. Example of acoustic labelling in Praat of the word [a'ju] (target word *iogurt* – 'yoghourt') produced by An at 1;5.

The annotation of prosodic prominence was conducted perceptually and at word-stress level. Catalan is a stress-accent language in which lexically stressed syllables generally serve as the main landing site for phrasal pitch accents (Prieto, Borràs-Comes, Cabré, Crespo-Sendra, Mascaró, Roseano, Sichel-Bazin & Vanrell, 2013). Word stress always hits one of the last three syllables of the morphological word. Prieto (2006) analyzed a corpus of adults addressing children and found that 35% of the words were monosyllables, 49% were disyllables and 13% were trisyllables. The remaining 3% of the data corresponded to longer words. Among the disyllabic forms, 63% were trochees and 37% iambs. Finally, among the trisyllabic forms, 72% were amphibracs. Importantly, no analysis of acoustic correlates of prominence such as duration or F0 tonal alignment was performed in our study because these correlates are still not stable at the ages in which children were analyzed (Astruc, Prieto, Payne, Post & Vanrell, 2013; Bonsdroff & Engstrand, 2005; Engstrand & Bonsdroff, 2004; Frota & Vigário, 2008; Vanrell, Prieto, Astruc, Payne & Post, 2011).

Figure 2 summarizes the procedure followed for speech coding.

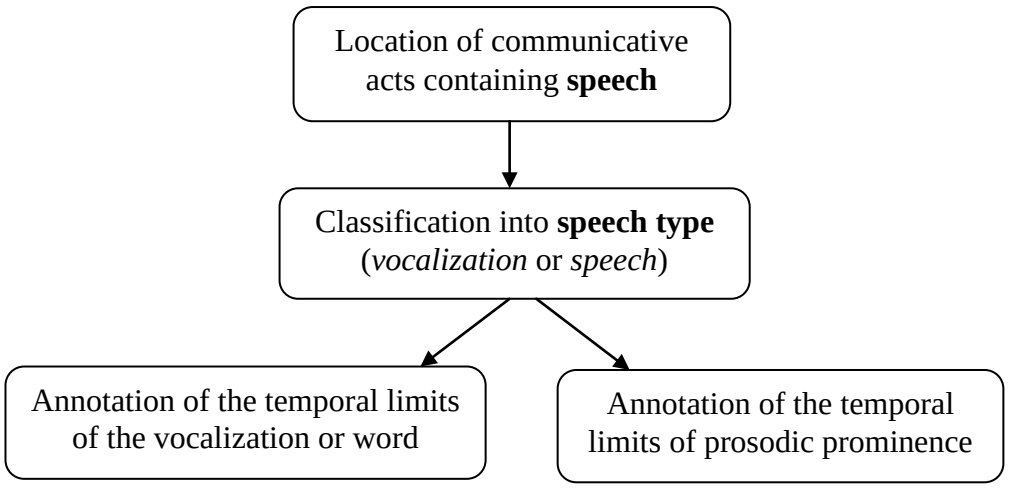

Figure 2. Summary of the steps followed during the speech coding.

2.3.2 Gesture coding

All communicative acts containing a gesture were coded using ELAN to determine their gesture type (tier 2 in figure 4). The following categories were taken into account (following Blake, Vitale, Osborne & Olshansky, 2005; Capone & McGregor, 2004; Ekman & Friesen, 1969; Iverson, Tencer, Lany & Goldin-Meadow, 2000): *pointing gesture*, a deictic gesture infants produce when extending the arm and the index finger towards an entity in order to direct the caregiver's attention to it; *reaching gesture*, a deictic gesture produced when the child extends the arm and opens the hand towards an entity in order to direct the caregiver's attention to it; *conventional gesture*, ritual actions such as head nodding to mean 'yes', head shaking to mean 'no', bye-bye gesture, clapping hands, kissing gesture, 'sh' gesture, and negating with the index-finger extended; *emotive gesture*, the child's expression of an emotional state, such as shaking arms when being angry, or shaking legs to protest, as opposed to the transmission of information; and finally *other gestures*, when the infant produced a proto-beat gesture, or an object-related action resembling an iconic gesture.

Next, all gestures classified as either pointing or reaching (i.e., which shared the feature of being deictic) were annotated regarding their motivation or intentionality. Gesture motivation was annotated in order to investigate potential influences of this factor on the temporal alignment of gesture and speech. Two categories were taken into account in this annotation, *imperative* or *declarative* (tier 3 in figure 4). A deictic gesture had an imperative motive if infants used it to ask the adult to retrieve an object for them, and a declarative motive if infants used it to share attention or inform the adult about

something. Most of the studies on pointing development support the dichotomy between imperative and declarative pointing gestures that was first proposed by Bates et al. (1975) and later corroborated by further research (Camaioni, Perucchini, Bellagamba & Colonnesi, 2004; Cochet & Vauclair, 2010; Liszkowski, 2007; Tomasello et al., 2007).[2]

To test the reliability of the gesture coding, the two independent coders that also participated in the previous reliability tests conducted two inter-transcriber reliability tests with a random subset of 20% of the data (145 cases), one in terms of gesture type and another in terms of gesture motive, in which all children and all ages were uniformly represented. In terms of gesture type (pointing, reaching, emotive, conventional, or others), overall agreement between coders was 95% and the kappa statistic value was 0.94, suggesting almost perfect agreement. For gesture motive (imperative or declarative), overall agreement was 86% and the kappa statistic value was 0.73, suggesting again a high degree of agreement between coders.

Finally, all gesture-speech combinations containing a pointing or reaching gesture were annotated in terms of their gesture phases (tier 4 in figure 4), following McNeill's (1992) and Kendon's (2004) observational measures: (a) the *preparation phase*, in which the arm moves from rest position until the stroke of the gesture; (b) the *stroke phase*, the interval of peak of effort in the gesture that expresses the meaning of the gesture (McNeill, 1992:83) or, in other words, the phase when the 'expression' of the gesture, whatever it may be, is accomplished and in which the movement dynamics of 'effort' and 'shape' are expressed with greater clarity (Kendon, 2004:112) (c) the

[2] The distinction between imperative and declarative has alternatives in the literature: Begus & Southgate (2012) and Southgate, van Maanen & Csibra (2007) propose that all infant pointing gestures have an interrogative function, Leavens (2009) state that they all have an instrumental function, and Moore and D'Entremont (2001) argue that all pointing gestures are motivated egocentrically.

retraction phase, in which the arm moves from the stroke position to rest position again. Additionally, another measure was annotated within the stroke of the gesture, namely the *gesture apex* (tier 5 in figure 4). Whereas the stroke of the gesture is an interval of time in the case of a deictic gesture during which the arm is maximally extended, the gesture apex is the specific point within the stroke interval at which the finger is maximally extended.

In order to locate the *stroke* and *apex* of the pointing gesture, we examined the video file (following Esteve-Gibert & Prieto, 2013; Levelt et al., 1985; Rusiewicz, 2010). ELAN allows precise navigation through the video recording, i.e., frame by frame. Though the software program can in principle permit an even more precise annotation (2 ms by 2 ms), this option could not be applied because the video was recorded at a frame rate of 25 frames per second. First, the stroke of the gesture was annotated in those video frames in which the arm was well extended with no blurring of the image, the fingertip being fully extended or not. Despite the absence of image blurring, the arm was not totally static during the interval of the gesture stroke, with the fingertip moving a few pixels back and forth. Next, the gesture apex was annotated in the specific video frame in which we located the furthest spatial excursion of the fingertip during the interval of time in which the arm was maximally extended (see still images at the bottom of figure 4). When infants performed the pointing gesture more slowly, this gesture peak could last more than one frame (normally two frames). In such cases, the gesture peak was considered to be the last of these video frames.

Figure 3 summarizes the procedure followed for gesture coding and shows an example of each type of gesture.

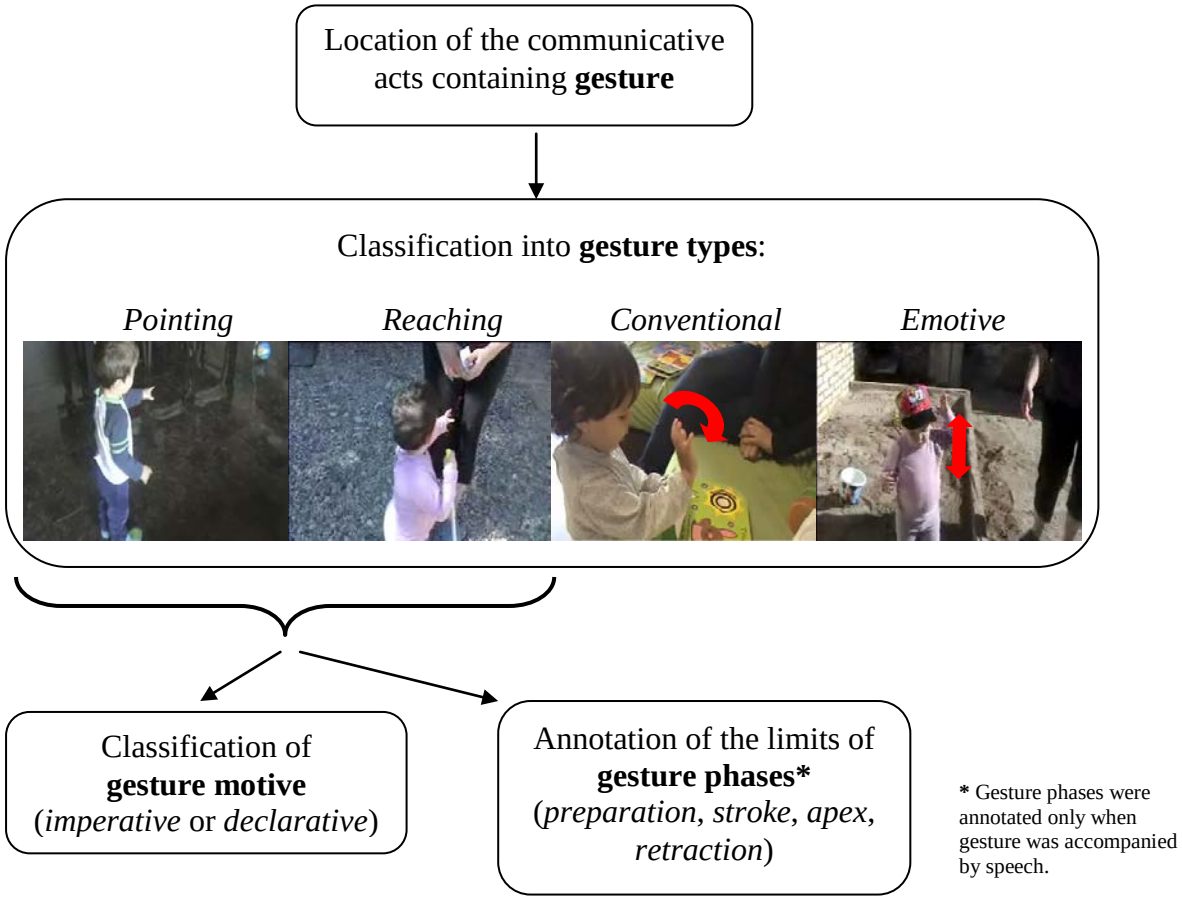

Figure 3. Summary of the gesture coding with a still image of every gesture type taken into account.

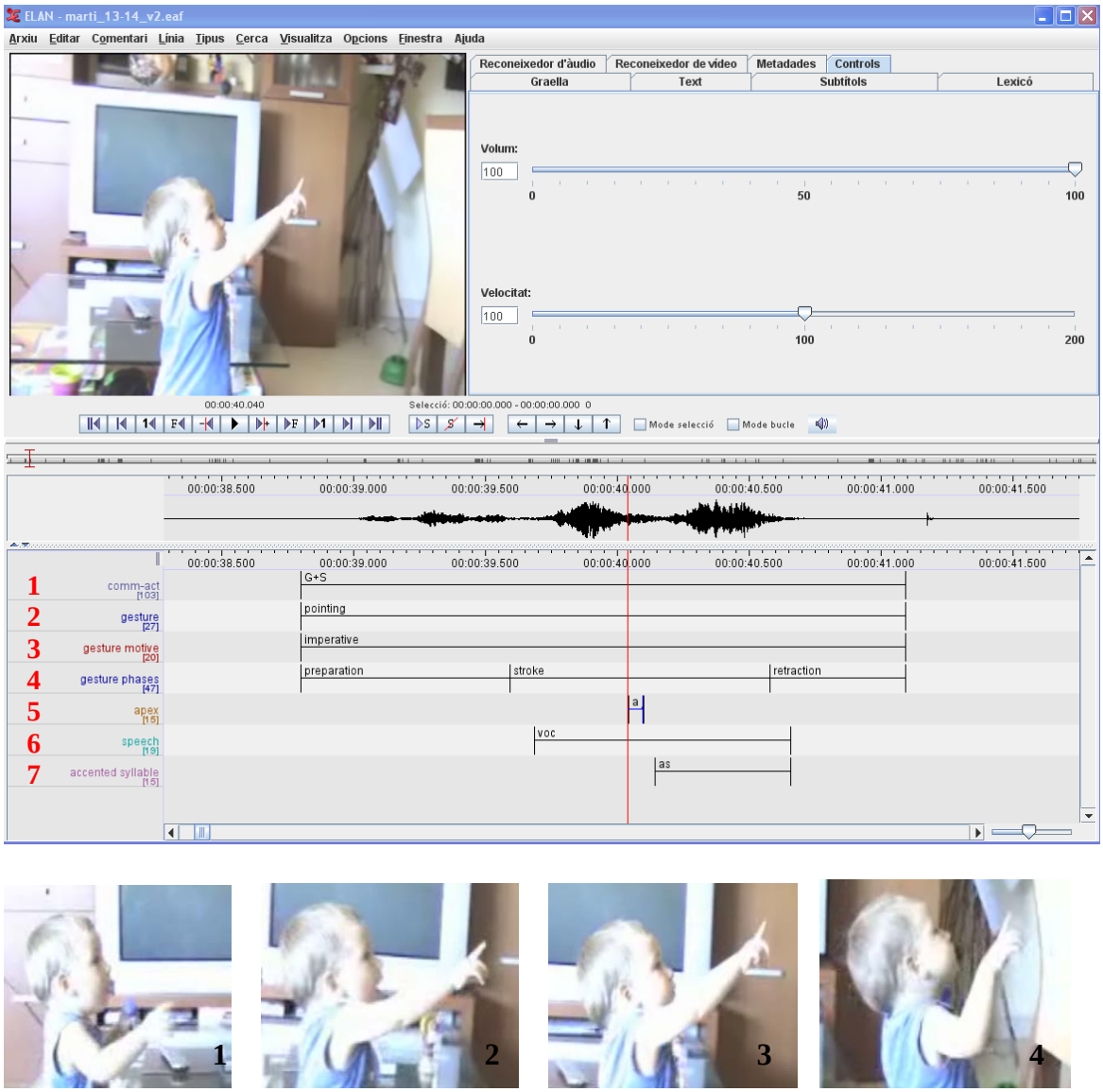

Figure 4. ELAN still image with all the annotated tiers (from 1 to 7). Lower frames, four specific enlarged images taken in the course of a pointing gesture which illustrate the four phases of a pointing gesture: (1) the preparation phase, (2) the stroke phase and before the apex is reached, (3) the apex, and (4) the retraction phase.

3. RESULTS

The main goal of this study is to analyze the early development of gesture and speech patterns, and it can be divided in two specific goals: (1) how do children combine gesture with speech across ages, and (2) how do they temporally align the two

modalities across age groups. Results are presented in two main sections, one for each aim.

3.1 How do children combine gesture with speech across ages?

In this section we explore how children combine gesture and speech across ages. Three main questions will be addressed: (1) When do children start producing most of their gestures in combination with speech? (2) In the first gesture-speech combinations, which gesture types do infants produce? And (3) which intentions do children convey in their first pointing gesture-speech combinations?

We first examined when children start combining gestures with speech. Figure 5 shows the distribution (expressed in percentages) of 'gesture-only' and 'gesture-speech combination' acts produced by infants across age groups (table 2 shows the raw numbers analyzed for each of the subgroups, including also the speech-only group). As can be observed in figure 5, of all the communicative acts containing gestures, at 0;11 most do not yet involve speech. At 1;1 infants produce roughly the same number of gestures accompanied by speech and gestures without speech, and from 1;3 onwards the proportion of gesture-speech combinations is higher than the proportion of gesture-only acts. Chi-squared tests of independence were carried out in order to test whether the ratio of 'gesture-only' to 'gesture-speech combination' acts differed across age groups. Results showed that the ratio of 'gesture-only' to 'gesture-speech combination' was statistically different at 0;11 (χ^2 (1, N=455) = 24.231, p< .001), at 1;3 (χ^2 (1, N=268) = 20.433, p< .001) and at 1;7 (χ^2 (1, N=166) = 38.554, p< .001), but not at 1;1 (χ^2 (1, N=259) = .004, p= .950) nor at 1;5 (χ^2 (1, N=249) = .486, p= .486). These results

indicate that at 0;11 children produce most of their gestures without accompanying speech, and that it is not until children are 1;3 that this tendency changes to such an extent that most of their gestures are produced together with speech, as seen in adults.

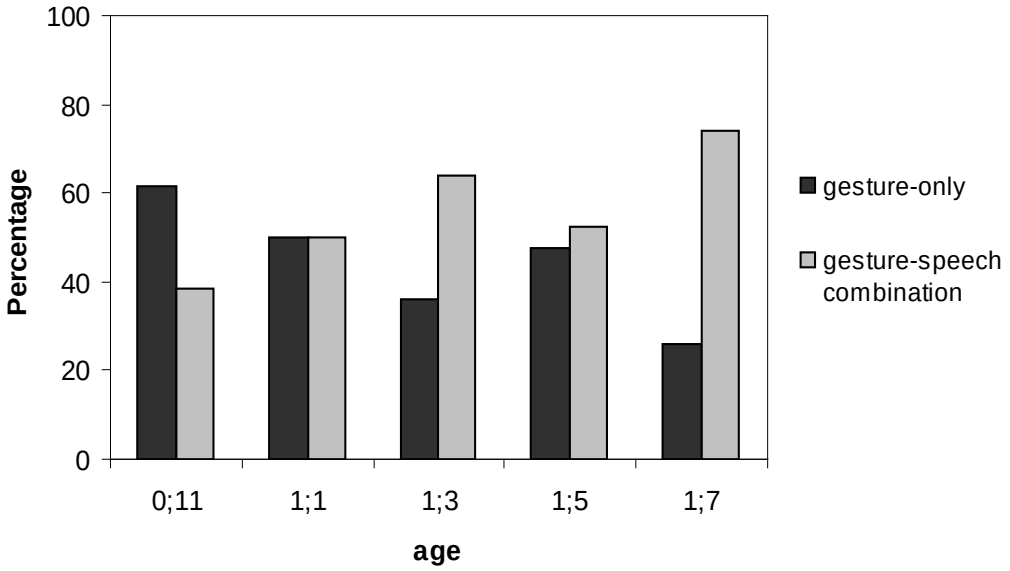

Figure 5. Ratio (expressed in percentages) of the distribution of all communicative acts containing gestures into the category of 'gesture-only' or 'gesture-speech combination'.

	Age	*Speech-only*	*Gesture-only*	*Gesture-speech combination*	*Total*
	0;11	710	280	175	1,165
	1;1	576	129	130	835
	1;3	722	97	171	990
	1;5	556	119	130	805
	1;7	546	43	123	712
Total		3,110	668	729	4,507

Table 2. Description of the data included in the analysis as a function of the ages and communicative act types.

With regard to the second research question, i.e., which gesture types do infants produce when they combine gesture with speech, table 3 shows the distribution (expressed in both absolute numbers and percentages) of gesture types across age groups in the gesture-speech combinations. The results show that deictic gestures were the most frequent gestures across all age groups. At 0;11 most of the gestures infants produced were deictic gestures (48.6%, divided into 34.9% pointing and 13.7% reaching), and emotive gestures (30.9%). At 1;1 the proportion of deictic gestures increased to 56.9% (40% pointing and 16.9% reaching) and emotive gestures represented 29.3% of the gestures. At 1;3 infants produced more deictic gestures than any other type of gesture (53.8% pointing and 20.5% reaching). At 1;5 conventional gestures increased compared to the previous ages but most gestures were nonetheless still deictic (40.8% pointing and 13.8% reaching). Finally, at 1;7 the proportion of pointing deictic gestures was higher than at all previous ages (65% pointing and 13.8% reaching).

	Pointing		*Reaching*		*Conventional*		*Emotive*		*Other*	
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
0;11	61	34.9	24	13.7	16	9.1	54	30.9	20	11.4
1;1	52	40.0	22	16.9	9	6.9	38	29.3	9	6.9
1;3	92	53.8	35	20.5	20	11.7	16	9.4	8	4.6
1;5	53	40.8	18	13.8	37	28.6	14	10.8	8	6.1
1;7	80	65.0	17	13.8	17	13.8	8	6.5	1	0.8

Table 3. Total numbers and percentages of gesture types across ages in gesture-speech combinations.

Regarding the gesture motives behind the pointing and reaching gestures, figure 6 (top panel) shows that infants produced a higher proportion of declarative deictic gestures than imperative deictic gestures at all ages. A look at the gesture types that infants used to convey these distinct gesture motives reveals interesting results: figure 6 (bottom left

panel) shows that when an imperative motive was conveyed, most times was by means of a reaching gesture, whereas figure 6 (bottom right panel) shows that when a declarative intention was conveyed, it was by means of a pointing gesture.

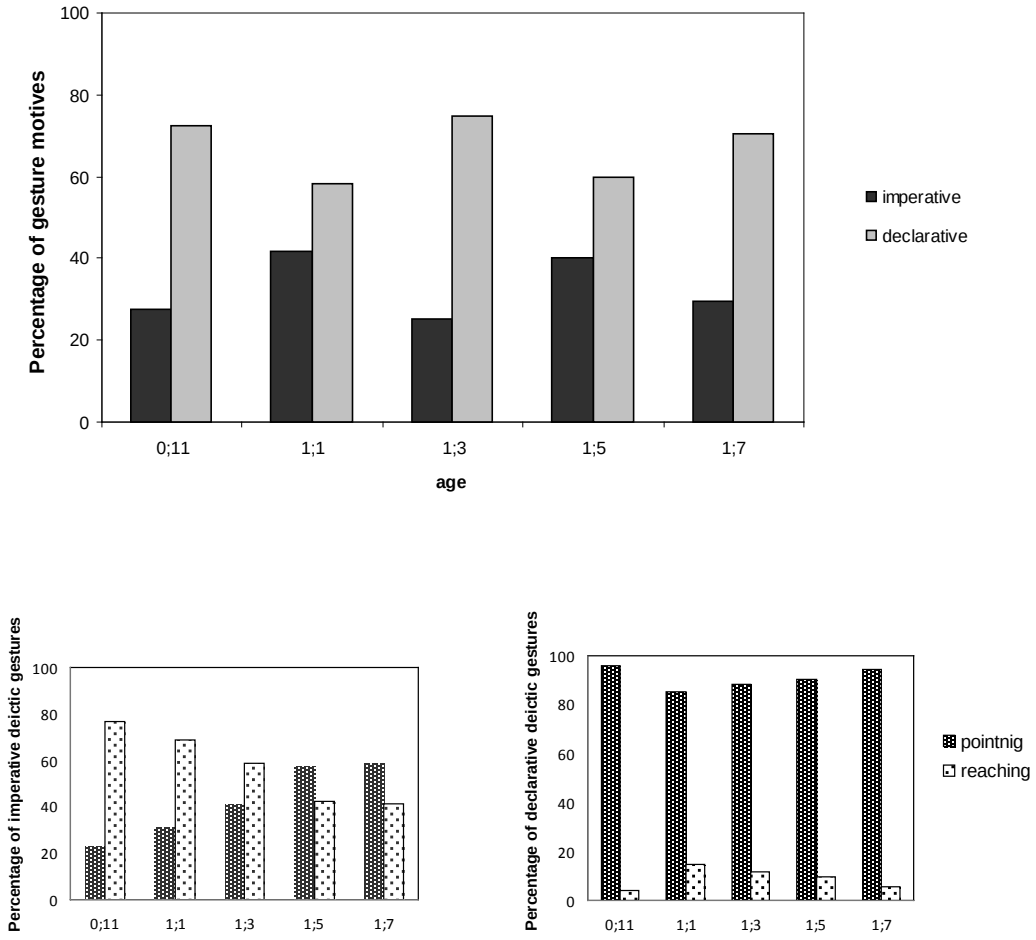

Figure 6. Relative proportions of imperative and declarative gestures across ages (top panel). Bottom panels, ratios of gesture types within imperative deictic gestures (left panel) and declarative deictic gestures (right panel).

The intention of deictic gestures was analyzed to investigate potential influences of this factor on the duration of the gesture and consequently on the temporal alignment of early gesture-speech combinations. Our hypothesis was that imperative pointing and reaching gestures would be longer, and thus would overlap more and be more easily

coordinated with speech. However, the statistical analyses revealed that no such effect occurred. Specifically, LMM analyses were carried out with stroke duration as the dependent variable, gesture motive (2 levels: imperative, declarative) as the fixed factor, and subject as a random factor. Results revealed no main effect of gesture motive on the duration of the gesture stroke ($F(1,4.47) = 1.453$, $p = .229$). For this reason gesture motive was not included as a fixed factor in any of the subsequent analyses of the temporal coordination of different gesture and speech landmarks.

3.2 How do children temporally coordinate gesture and speech across ages?

The main goal of this section is to assess how infants temporally coordinate deictic gesture and speech combinations across the ages analyzed. Following the adult studies on gesture-speech temporal coordination, the synchronization between deictic gesture-speech combinations was analyzed at four different levels: (1) the temporal relationship between the gesture onset and the onset of the vocalization, to compare our results in children with studies for adults suggesting that the onset of the gesture always precedes the onset of speech (Butterworth & Beattie, 1978; Ferré, 2010; Levelt et al., 1985); (2) the temporal coordination between the stroke onset and the onset of speech, to compare our results in children with results for adults suggesting that the stroke of the gesture coordinates with the onset of speech (Bergmann et al., 2011; Ferré, 2010; Roustan & Dohen, 2010); (3) the coordination between the stroke onset and the onset of the accented syllable in speech, to compare our results with results for adults suggesting that gesture strokes are coordinated with prosodically prominent syllables (Krahmer & Swerts, 2007; Leonard & Cummins, 2010; Loehr 2007); and (4) the coordination

between the gesture apex and the end of the accented syllable in speech, to compare our results in children with studies on adult coordination suggesting that the gesture apex occurs within the limits of the prosodically prominent syllables (De Ruiter, 1998, Esteve-Gibert & Prieto, 2013; Rochet-Capellan, Laboissière, Galván & Schwartz, 2008).

All statistical analyses in this section were performed by applying a linear mixed model (LMM; West, Welch & Galecki, 2007) using SPSS Statistics 16.0 (SPSS Inc., Chicago IL). West et al. (2007), and Baayen, Davidson and Bates (2008) state that LMMs are the appropriate model for analyzing unbalanced longitudinal data, since they allow for subjects with missing time points (i.e., unequal measurements over time for individuals), have the capacity to include all observations available or all individuals in the analysis and cope with missing data at random. As the authors point out, linear mixed models can accommodate all of the data that are available for a given subject, without dropping any of the data collected from that subject.

First, the coordination between the gesture onset and the onset of the associated speech was analyzed. The temporal distance between the onset of the vocalization and the gesture onset was the dependent variable, age was the fixed factor (5 levels: 0;11, 1;1, 1;3, 1;5 and 1;7), and subject was included as a random factor. The analysis revealed a statistically significant effect of age on the distance between gesture onset and onset of speech ($F(4,462.530) = 9.998$, $p < .001$) (see table 4 for statistic coefficients). LSD pair-wise comparisons revealed that the mean distance between gesture onset and onset of speech varied significantly between 0;11 and 1;1 ($p < .05$), between 0;11 and 1;3 ($p < .001$), between 0;11 and 1;5 ($p < .001$), between 0;11 and 1;7 ($p < .001$), between 1;1

and 1;7 ($p < .01$), and between 1;3 and 1;7 ($p < .05$). Figure 7 shows the box plots of this measure across ages, showing that the gesture onset precedes the onset of speech at all ages, but that the tendency is for this distance to decrease as the child grows up and for the variance in this distance measure to decrease as well. In adult studies, it has been found that gesture onset precedes speech onset. For example, Butterworth and Beattie (1978) suggested that 'gesture onset preceded voice onset by an interval whose magnitude was inversely related to the lexical affiliate's rated familiarity', and Ferré (2010) found that 70% of gestures start before the onset of the corresponding intonational phrase. None of these studies, however, detail the time lag they found between the onset of gesture and the onset of speech. Our results thus show that infants coordinate these two landmarks in an adult-like way in the sense that gesture onset occurs before speech onset.

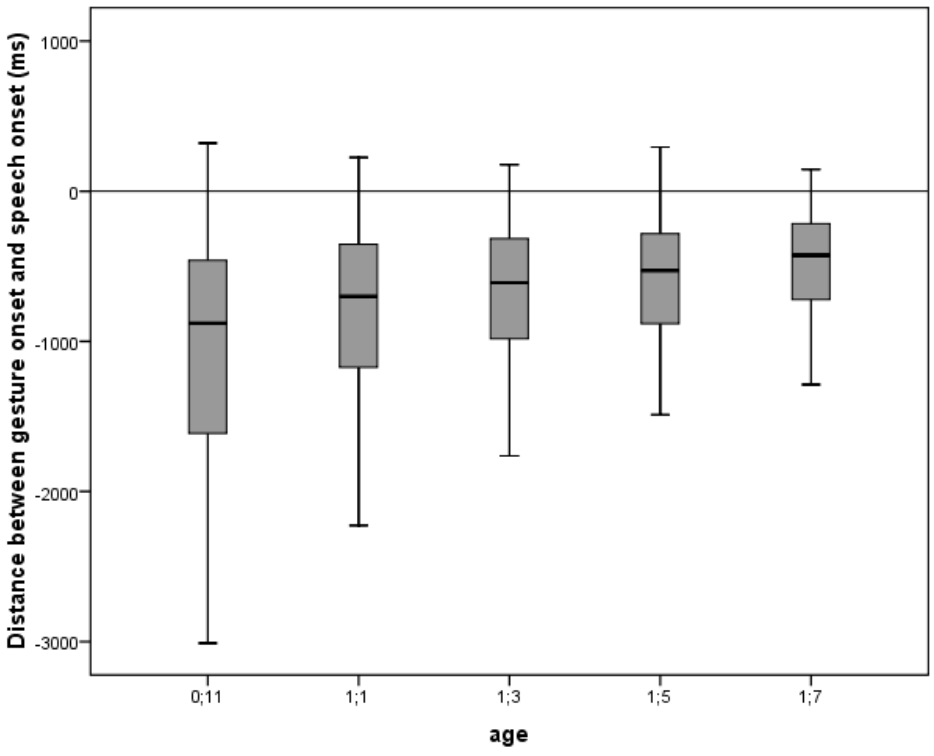

Figure 7. Temporal distance (in milliseconds) between gesture onset and the onset of speech across ages. Positive values (> 0) represent those cases in which the gesture onset occurs after

the onset of speech, while negative values (< 0) represent cases in which the gesture onset occurs before the onset of speech.

Dependent variable: Temporal distance between gesture onset and onset of speech					
	Estimates	*Residual variance*	*Standard deviation*	*F value*	*P value*
Fixed factor (age)	-505.53	-	756.88	9.998	.000***
Random factor (subject)	-	523063.03	723.23	-	-

Table 4. Table of coefficients with estimates, standard deviation, F value and p value for the fixed factor, and residual variance and standard deviation for the random factor[3].

Second, following studies with adults (Bergmann et al., 2011; Ferré, 2010; Roustan & Dohen, 2010) we introduced a more fine-grained coordination measure which takes into account the temporal position of the most prominent period in the deictic gesture, namely the stroke. Thus, the coordination between the stroke onset and the onset of speech was analyzed. The temporal distance between the onset of the stroke and the onset of speech was the dependent variable, age was the fixed factor (5 levels: 0;11, 1;1, 1;3, 1;5 and 1;7), and subject was included as a random factor. The statistical analysis indicated that age did not have an effect on this distance ($F(4,434.639) = 2.066$, p =

[3] The SPSS formula used to calculate these values was the following:

```
MIXED
  distance_gesture_onset_and_speech_onset  BY age
  /CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR
  (0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE)
  PCONVERGE
  (0.000001, ABSOLUTE)
  /FIXED = age  | SSTYPE(3)
  /METHOD = REML
  /PRINT = CPS COVB DESCRIPTIVES G SOLUTION TESTCOV
  /RANDOM INTERCEPT  | SUBJECT(subject) COVTYPE(VC)
/EMMEANS = TABLES(age) COMPARE ADJ(LSD).
```

All the subsequent LMM analyses included in the study were performed by applying this SPSS formula, with the only difference of substituting the dependent variable "distance_gesture_onset_and_speech_onset" with the dependent variable relevant in each particular analysis.

.084) (see table 5 for statistic coefficients). Figure 8 displays the distance between stroke onset and onset of speech across ages. The figure shows that there is a close temporal alignment between the two modalities across ages, as found in studies of adults. Though the age factor was not significant, a slight tendency is observed in figure 8 showing that during the babbling stage (0;11 and 1;1), the stroke onset is aligned with the onset of speech, and that as the children's linguistic abilities develop (1;3 onwards), the speech tends to start slightly before the stroke. Studies in adult gesture-speech coordination which took into account these two measures found that 72% of the gesture strokes start before the onset of speech (Ferré, 2010), that stroke onset precedes speech onset on average by 129.89 milliseconds (Bergmann et al., 2011) and that the maximum extension of the finger occurs within or close to the focus constituent (Roustan & Dohen, 2010).

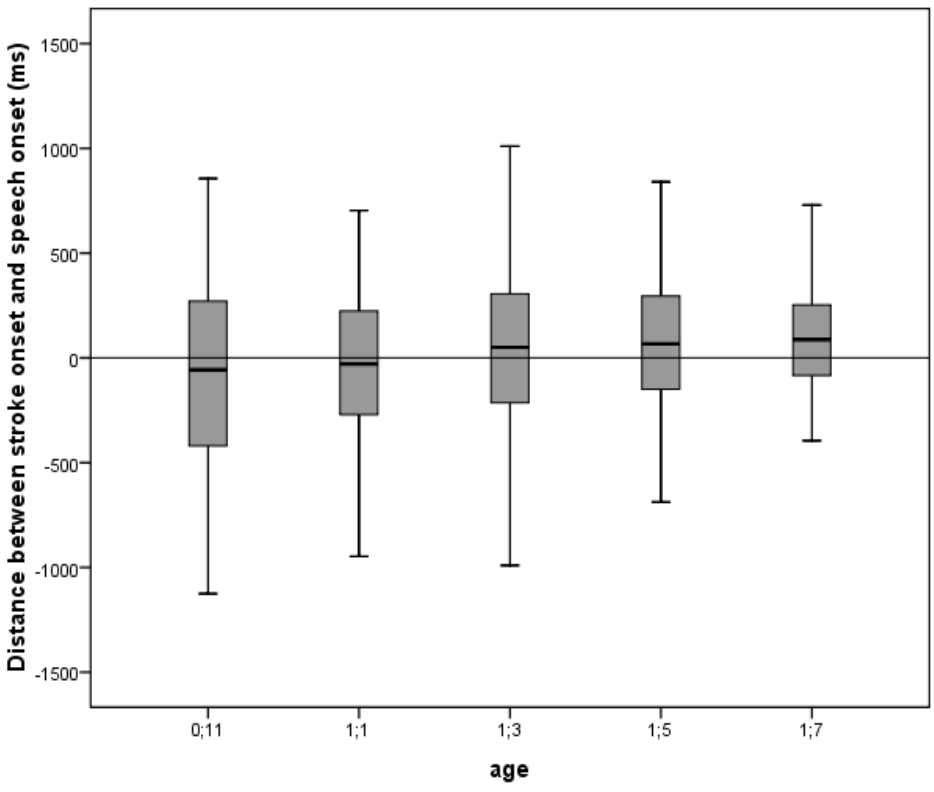

Figure 8. Distance between stroke onset and onset of speech (in milliseconds) across ages. Positive values (> 0) represent those cases in which stroke onset occurs after speech onset, while negative values (< 0) represent cases in which stroke onset occurs before speech onset.

Dependent variable: Distance between stroke onset and onset of speech (in ms)					
	Estimates	*Residual variance*	*Standard deviation*	*F value*	*P value*
Fixed factor (age)	79.31	-	3123.93	2.066	.084
Random factor (subject)	-	157430.99	396.77	-	-

Table 5. Table of coefficients with estimates, standard deviation, F value and p value for the fixed factor, and residual variance and standard deviation for the random factor.

Third, we introduced the location of the prosodically prominent syllables in the coordination analysis, as adult studies have highlighted the importance of this anchoring site for deictic gestures (Krahmer & Swerts, 2007; Leonard & Cummins, 2010; Loehr 2007). The coordination between the stroke onset and the onset of accented syllable was analyzed. We expected an even tighter coordination between the start of the gesture stroke and the start of the accented syllable as children develop, since studies with adults have reported a close synchronization between these two landmarks. Thus, the dependent variable was the distance between stroke onset and the onset of the accented syllable, the fixed factor was age (5 levels: 0;11, 1;1, 1;3, 1;5 and 1;7), and subject was introduced as a random factor. The statistical analysis indicated that age did not have an effect on the distance between the stroke onset and the onset of accented syllable ($F(4,431.213) = .595$, $p = .667$) (see table 6 for statistic coefficients). Figure 9 displays the distance between onset of the stroke and the onset of the accented syllable across ages. Though age does not significantly affect this measure, a tendency can be observed: at the babbling stage (at 0;11 and 1;1), the stroke tends to slightly precede the onset of the accented syllable, whereas from the onset of word production onwards

(from 1;3 to 1;7), the stroke onset and the onset of the accented syllable are very closely aligned, and less data variance is observed. These results are in accordance with adult studies showing a co-occurrence between those landmarks (Krahmer & Swerts, 2007; Leonard & Cummins, 2010; Loehr 2007).

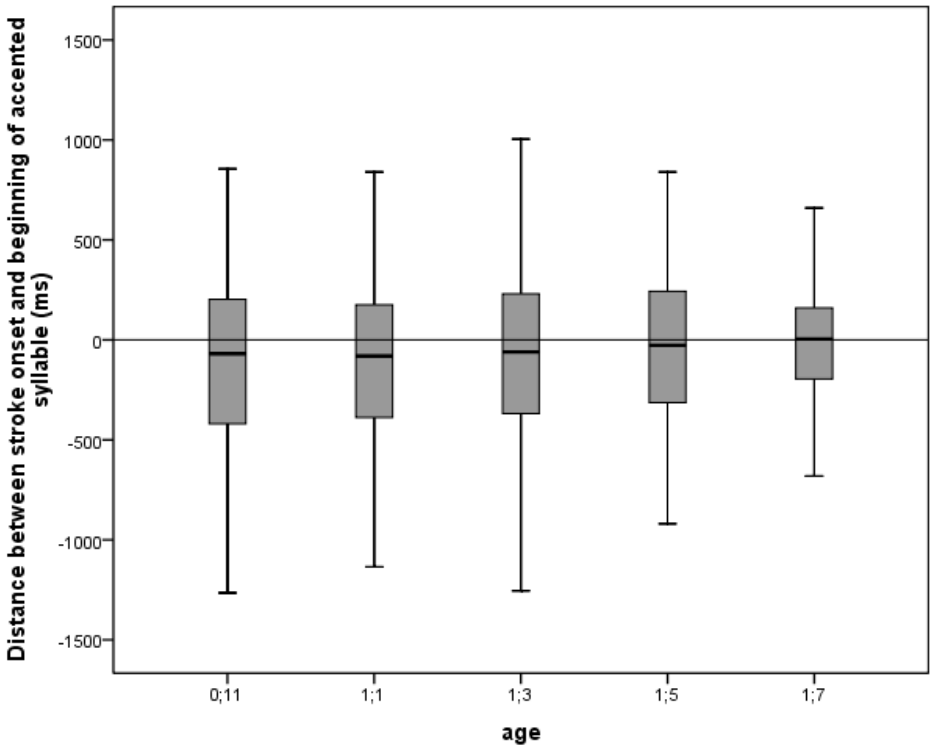

Figure 9. Distance between stroke onset and onset of the accented syllable across ages (in milliseconds). Positive values (> 0) represent those cases in which stroke onset occurs after onset of accented syllable, while negative values (< 0) represent cases in which stroke onset occurs before onset of accented syllable.

Dependent variable: Distance between stroke onset and onset of accented syllable (in ms)					
	Estimates	*Residual variance*	*Standard deviation*	*F value*	*P value*
Fixed factor (age)	-1.004	-	412.383	.595	.667
Random factor (subject)	-	168135.28	410.043	-	-

Table 6. Table of coefficients with estimates, standard deviation, F value and p value for the fixed factor, and residual variance and standard deviation for the random factor.

Fourth, the temporal distance between the gesture apex and the end of the accented syllable was also analyzed. The dependent variable was the distance between the gesture apex and the end of the accented syllable, the fixed factor was age (5 levels: 0;11, 1;1, 1;3, 1;5 and 1;7), and subject was the random factor. The statistical analysis showed that age did not have a main effect on the position of the gesture apex within the accented syllable ($F(4,439.933) = 1.127$, $p = .343$) (see table 7 for statistic coefficients). Thus, at all ages the gesture apex precedes the end of the accented syllable (*see* figure 10), but different tendencies can be observed: during the babbling stage, the apex tends to be located further from the end of the accented syllable, and the box plots show higher variation; as the children's linguistic abilities develop, however, the gesture apex tends to occur closer to the end of the accented syllable and the variation diminishes. Studies on the coordination in adults between gesture apex and the accented syllable showed that the gesture apex occurs between 350 and 0 ms prior to the end of the accented syllable (De Ruiter, 1998; Esteve-Gibert & Prieto, in press), so our results show that infants coordinate these two landmarks in an adult-like way at the babbling stage in the sense that they produce the gesture apex before the accented syllable is finished.

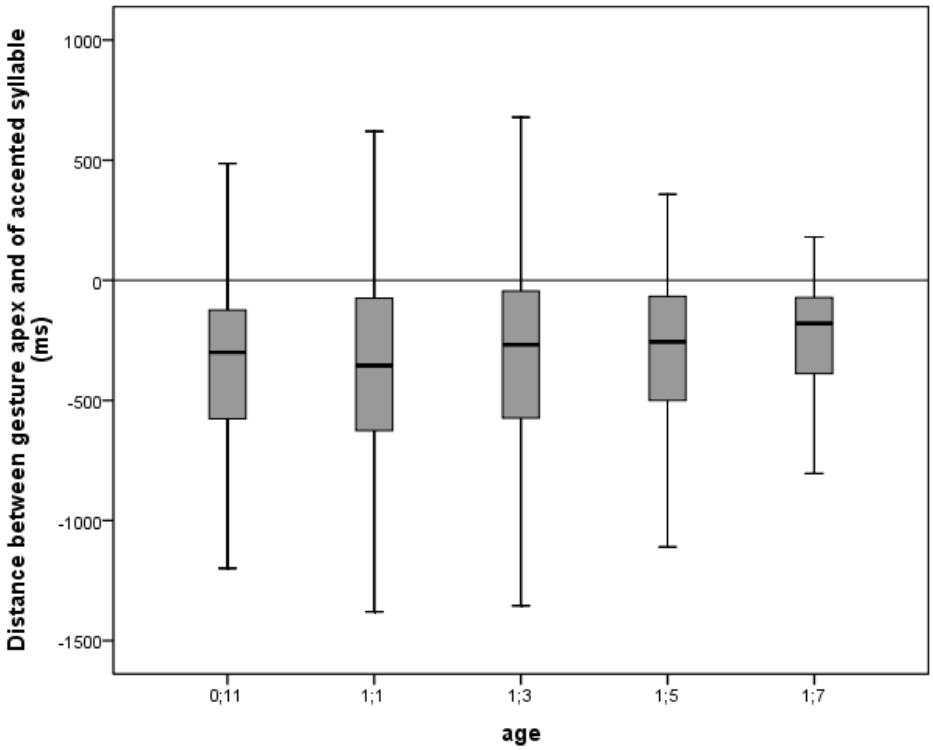

Figure 10. Distance between the gesture apex and the end of the accented syllable (in milliseconds) across ages and as a function of the metrical patterns. Positive values (> 0) represent those cases in which the gesture apex occurs after the end of the accented syllable, and negative values (< 0) are those cases in which the gesture apex occurs before the end of the accented syllable.

Dependent variable: Distance between gesture apex and end of accented syllable (in ms)					
	Estimates	*Residual variance*	*Standard deviation*	*F value*	*P value*
Fixed factor (age)	-221.52	-	402.752	1.127	.343
Random factor (subject)	-	153665.47	392.002	-	-

Table 7. Table of coefficients with estimates, standard deviation, F value and p value for the fixed factor, and residual variance and standard deviation for the random factor.

Summarizing, children show an adult-like pattern of coordination between the distinct gesture and prosodic landmarks analyzed across ages. First, gesture onset precedes speech onset (figure 7), although there are significant differences across ages: at the

babbling stage, the distance between the two points is higher and there is more variation in the data, whereas at the single-word period this distance is more similar to what previous studies have found in adult data, and the variation is significantly reduced. Second, gesture and speech are very tightly coordinated when landmarks take into account prominence: infants align more closely the beginning of the gesture stroke with the speech onset, and also the beginning of the stroke with the beginning of the accented syllable. Importantly, age was not a significant main factor in either of the two coordination analyses, yet figures 8 and 9 show that the absolute distance in time and variance across age groups is smaller in older children. Third, the temporal distance between gesture apex and the end of the accented syllable shows adult-like patterns in the sense that this gesture apex precedes the end of the accented syllable (figure 10). Although this coordination does not vary significantly across ages, a finer coordination is again observed at the single-word period. Altogether, these findings suggest clear adult-like patterns of gesture-speech integration in the very first multimodal utterances produced by the children already at the babbling stage.

4. DISCUSSION

This study explored the patterns of gesture and speech combinations in infants from the babbling to the single-word period, as well as the temporal alignment between the two modalities when they are combined. The analysis of approximately 24 hours of naturalistic recordings of four Catalan children in five consecutive developmental stages, namely 0;11, 1;1, 1;3, 1;5 and 1;7, provided us a total of 4,507 communicative acts. An infant's act was considered to be intentional if (a) the coder perceived or

judged the infant's act as based on awareness and deliberate execution, if (b) infants produced it in a joint attention frame, or if (c) the parental reactions before or after the acts suggested so. While these measures are not totally objective, they have been proven to be a reliable measure when correlating the adults' inclination to interpret infants' acts as intentional and the infants' later development of cognitive capacities (Olson, Bates & Bayles, 1982; Sperry & Symons, 2003). In the present study, two independent coders performed an inter-transcriber reliability test by identifying the communicative acts from 10% of the observational data. Results of this analysis reflect that although this method resulted into some coding errors, there was still substantial agreement between coders when locating communicative acts (the overall agreement was 83% and the free marginal kappa statistic was 0.67).

Summarizing, three main results can be highlighted from the data: first, it is from the onset of word production that gesture starts to be produced mainly in combination with speech with an intentional purpose; second, in these early gesture-speech combinations most of the gestures are deictic gestures (pointing and reaching) with a declarative communicative purpose; and third, there is clear evidence of temporal coordination between deictic gesture and speech already at the babbling stage in the sense that (a) gesture onset starts before speech onset, (b) the stroke onset is temporally coordinated with the onset of speech and with the onset of the prominent syllable, and (c) the gesture apex occurs before the end of the accented syllable. In the following paragraphs we discuss one by one these results in more detail, reporting the main findings suggested by our statistical analyses. Although only four infants were analyzed in this study, we believe that the large amount of data obtained can compensate for the small number of subjects.

The results of the longitudinal analysis show that it is not until the onset of word production that infants combine communicative gesture and speech significantly more often than producing gesture-only acts. Our results show that at 0;11 children produce more gesture-only acts than gesture-speech combinations. At 1;1 the proportion of gesture-speech combinations is still not higher than the proportion of gesture-only acts, though they have increased with respect to the previous ages analyzed. However, from 1;3 onwards children start producing their first words and also start combining gesture with speech significantly more often than producing gesture-only acts. These results confirm those by Butcher and Goldin-Meadow (2000). The authors started their analysis when children began producing their first words and ended it when they produced two-word combinations and found that children started combining gesture with speech at the transition between the one-word and two-word periods. Our study has enlarged the window analyzed by Butcher and Goldin-Meadow (2000), by focusing on the development of multimodality at the transition from the babbling period to the single-word period. All in all, our findings support those of Butcher and Goldin-Meadow (2000) with respect to an early infants' ability to combine gesture and speech.

A second finding of the study is that deictic gestures (pointing and reaching) are the most frequent gesture-speech combinations in the age range analyzed, and that at all ages children produced more deictic gestures with a declarative purpose than with an imperative purpose. Interestingly, we observe that imperative deictic gestures mostly take the form of reaching gestures, whereas declarative deictic gestures almost always take the form of pointing gestures, corroborating previous studies in the field (Blake, O'Rourke & Borzellino, 1994; Cochet & Vauclair, 2010; Colonnesi, Stams, Koster &

Noom, 2010; Leung & Rheingold, 1981). However, our results could be biased by the specific contexts in which these children were recorded, as they were recorded during free-play sessions at their homes and while interacting with their parents, and these play situations might not reflect the total output of gestures infants produce. For instance, it might well be that in an eating situation infants produce a higher proportion of imperative pointing gestures compared to declarative ones. Importantly, it should be borne in mind that our analysis of the development of the gesture motives and gesture types constitutes a description of the corpus under analysis and no strong conclusions can be drawn from it.

An important focus of our investigation was to describe developmental patterns related to how early gesture and speech combinations are temporally coordinated. In order to investigate this, we analyzed in detail the temporal coordination between gesture and speech by taking into account specific measurements that previous studies with adults had proposed to be crucial in characterizing gesture-speech alignment. Studies in adult multimodal communication have shown that gesture onset precedes speech onset (Butterworth & Beattie, 1978; Ferré, 2010; Levelt et al., 1985), that the gesture stroke slightly precedes the onset of speech (Bergmann et al., 2011; Ferré, 2010; Roustan & Dohen, 2010), that the most prominent part of the gesture coincides with the most prominent part of speech (Krahmer & Swerts, 2007; Leonard & Cummins, 2010; Loehr, 2007), and that the gesture apex is produced within the limits of the prominent syllable in speech (De Ruiter, 1998, Esteve-Gibert & Prieto, in press; Rochet-Capellan et al., 2008). Our analysis reveals that infants' behaviour shows all these alignments even before they can produce their first words.

First, our results on the distance between gesture onset and onset of speech reveal that infants at 0;11 show the adult-like pattern in that gesture starts before speech. However, our statistical analysis also revealed that age significantly affects this measure because in the babbling period the distance between the two measures is significantly higher than in the single-word period.

Second, our results show that already at 0;11 and across all the stages analyzed infants coordinate the gesture stroke with the onset of speech because they produce these two landmarks simultaneously. Some studies with adults have shown that the stroke onset precedes the speech onset (Bergmann et al., 2011; Ferré, 2010; Roustan & Dohen, 2010), while some others found that the gesture stroke starts when the accented syllable has already been initiated in monosyllables and trochees, and that both landmarks occur almost simultaneously in the iambic condition (Esteve-Gibert & Prieto, in press). These different findings with adults could be due to the fact that they analyzed different types of gesture: those who found that the gesture stroke precedes the speech onset had analyzed a mixture of iconic, deictic, and discourse gestures, and those who found that the gesture stroke follows the speech onset had analyzed only deictic gestures. In the present study, infants produce much simpler speech structures at the ages analyzed, mostly monosyllables and a few disyllables, and this might influence the coordination between stroke onset and onset of speech. Importantly, our statistical analysis revealed no effect of age on the distance between stroke onset and onset of speech, contrary to what Butcher and Goldin-Meadow (2000) found in their study. These authors found that the coordination between these two measures was significantly affected by age, adult-like synchronization being present at the two-word stage but not at the single-word period. The difference in results might be due to the specific anchoring points taken into

account: Butcher and Goldin-Meadow (2000) considered gesture-speech combinations to be synchronous if the vocalization occurred on the stroke of the gesture or at the peak of the gesture, whereas in the present study we took into account the onset of both the gesture stroke and speech.

Third, the tightest adult-like coordination between gesture and speech is observed when prominence in gesture and prominence in speech are taken into account: infants produce the stroke onset coinciding with the onset of the accented syllable, just as adults do. Crucially, our statistical analysis revealed that this finding is not significantly affected by age. It is interesting to note in figures 8 and 9 that infants temporally align the gesture stroke as closely with the onset of the accented syllable as they do with the speech onset. This fact might be because these two measures in speech are very close to each other given that most of the infants' first vocalizations and words are monosyllables or have word-initial stress.

Fourth, our analysis of the alignment between gesture apex and the end of the accented syllable revealed adult-like patterns already at the babbling stage to the extent that infants produce the gesture apex before the end of the accented syllable. Studies on adult coordination of gesture and speech have shown that gesture apex occurs between 350 ms and 0 ms prior to the end of the accented syllable (De Ruiter, 1998; Esteve-Gibert & Prieto, in press; Rochet-Capellan et al., 2008), and our results demonstrate that infants already show signs of the same behaviour before they produce their first words.

All in all, our results show adult-like patterns when children combine and synchronize gesture and speech at an early age in language development and before they are able to produce two-word combinations. Thus, already at the babbling stage infants produce the

gesture onset before the onset of the corresponding speech (and this alignment is finely tuned at the single-word stage), they temporally coordinate the gesture stroke, i.e., the most prominent part of the gesture, with the speech onset and, crucially, with the onset of the accented syllable, and they produce the gesture apex before the end of the accented syllable.

These results expand on those reported in Butcher and Goldin-Meadow (2000) in two ways: first, our study focuses on children at the transition from the babbling to the single-word period and not at the transition between one- and two- word combinations, because it is at this earlier age that infants start producing gestures like pointing or reaching with a communicative purpose; and second, our study makes a detailed analysis of the temporal coordination between the two modalities based on the latest results in the field. Specifically, three main acquisition results can be highlighted in three points: (1) children start producing most of their gestures in combination with speech in an adult-like fashion at the early single-word period; (2) in these early combinations gesture onset always precedes speech onset, and this alignment is finely tuned at the single-word period compared to the babbling stage; and (3) both modalities are temporally coordinated in an adult-like way when gesture and acoustic prominences are taken into account: the stroke onset co-occurs with speech onset at all ages, the stroke onset precedes the beginning of the accented syllable at all ages, and the gesture apex is located before the end of the accented syllable at all ages. These results suggest that infants coordinate gesture and prosodic structures already before they produce their first words.

5. CONCLUSIONS

This study was intended to contribute to the body of research on the development of gesture and speech combinations in infants. Previous studies based on the dynamic systems theory (Iverson & Thelen, 1999) suggest that the early coordination between rhythmic motor and vocal movements is a precursor to the adult system in which gesture and speech are combined (Ejiri & Masataka, 2001; Iverson & Fagan, 2004). However, few studies have investigated the specific patterns of the early coordination of communicative gestures with speech (nor are there many on early rhythmic movements). To our knowledge, only Butcher and Goldin-Meadow (2000) have previously explored the question of when children learn to combine the two modalities and the way they synchronize them. They analyzed children at the transition between the one- and two-word stages and found that infants started combining the two modalities at the single-word period and started synchronizing them in an adult way at the two-word stage. Our results on the transition between the babbling stage and the single-word period confirm those by Butcher and Goldin-Meadow (2000) in the sense that it is at the single-word period that both modalities start being combined. Also, our study extends the work by Butcher and Goldin-Meadow (2000) because we analyze infants already from the babbling stage and because we examine in more detail the temporal coordination of deictic gesture-speech combinations.

In this respect, and based on recent findings in the literature on adult temporal coordination of gesture and speech, our analyses include coordination measurements related to prosodic and gestural prominence that have been found to play a crucial role in the temporal coordination between gesture and speech. And our results show that

there is evidence of temporal coordination between communicative gesture and speech already at the babbling stage because gestures start before their corresponding speech, because the stroke onset coincides with the onset of the prominent syllable in speech, and because the gesture apex is produced before the end of the accented syllable.

Various models of gesture and speech production have investigated the relation between gestures and speech. Theoretical models of gesture production such as the 'Growth Point Theory' by McNeill (1992), the 'Tradeoff Hypothesis' by Bangerter (2004), De Ruiter (2006), and Melinger and Levelt (2004), the 'Lexical Access Hypothesis' by Krauss, Chen and Gottesman (2000), or the 'Information Packaging Hypothesis' by Kita (2000) all try to account for the strong interrelation and influence between gesture and speech that characterize human communication. These models differ significantly regarding the semantic role of gestures with respect to speech and vice-versa, or the phases in which gestures are conceptualized, planned, and executed. However, they all agree on the close temporal integration of gesture and speech in production. We believe that the present findings provide evidence for this integration, and from a developmental point of view.

Indeed, our results suggest that there is a temporal coordination of communicative gesture and speech from the very first stages of language production. Yet strong claims about these integration patterns can only be made after more data is analyzed. Our study has been limited to the analysis of 4,507 longitudinal observations of naturalistic interactions between four infants and their caregivers. We think that a larger number of subjects should be analyzed in the future and that more experimental data in a controlled setting (with higher audio quality and movement trackers) will be useful to provide

more solid confirmation for our claim that gesture and speech form a temporally integrated system from the onset of language production.

5. ACKNOWLEDGEMENTS

A preliminary version of this paper was presented at the conferences *Architectures and Mechanisms for Language Processing* (Paris, France, September 1-3, 2011) and *Gesture and Speech in Interaction* (Bielefeld, Germany, September 5-7, 2011). We would like to thank participants at those meetings, especially Marc Swerts, Heather Rusiewicz, and Stefan Kopp, for their helpful comments. We thank Aurora Bel, Louise McNally, and José Ignacio Hualde for being part of the PhD project defense of the first author of the study and for their helpful observations. We also thank Ulf Liszkowski for his useful comments on the analysis of the intentions behind pointing gestures, Simone Bijvoet and Joan Borràs for their help with the statistical analysis, and Alfonso Igualada and Santiago González-Fuente for the inter-rater reliability tests. Finally, we are really grateful to the children and the children's parents for voluntarily taking part in this study.

This research has been funded by two research grants awarded by the Spanish Ministry of Science and Innovation (FFI2009-07648/FILO "The role of tonal scaling and tonal alignment in distinguishing intonational categories in Catalan and Spanish", and FFI2012-31995 "Gestures, prosody and linguistic structure"), by a grant awarded by the Generalitat de Catalunya ((2009SGR-701) to the *Grup d'Estudis de Prosòdia*, and by the Consolider-Ingenio 2010 (CSD2007-00012) grant.

6. REFERENCES

Astruc, L., Payne, E., Post, B., Vanrell, M.M., Prieto, P, 2013. Tonal targets in early child English, Spanish, and Catalan. Lang. Speech 56(2).

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59, 390–412.

Bangerter, A., 2004. Using pointing and describing to achieve joint focus of attention in dialogue. Psychol. Sci. 15, 415–419.

Bates, E., Camaioni, L., Volterra, V., 1975. The acquisition of performatives prior to speech. Merrill Palmer Q. 21, 205–224.

Bergmann, K., Aksu, V., Kopp, S., 2011. The Relation of Speech and Gestures: Temporal Synchrony Follows Semantic Synchrony. In: Proceedings of the 2nd Workshop on Gesture and Speech in Interaction, Bielefeld.

Blake, J., O'Rourke, P., Borzellino, G., 1994. Form and function in the development of pointing and reaching gestures. Infant Behav. Dev. 17, 195–203.

Blake, J., Vitale, G., Osborne, P., Olshansky, E., 2005. A cross-cultural comparison of communicative gestures in human infants during the transition to language. Gesture 5, 201–217.

Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer [Computer program]. Version 5.3.04, retrieved 12 January 2012 from http://www.praat.org/.

Bonsdroff, L., Engstrand, O., 2005. Durational patterns produced by Swedish and American 18- and 24-month-olds: Implications for the acquisition of the quantity contrast. In: Papers from the 18th Swedish Phonetics Conference, Department of Linguistics, Gothenburg, pp. 59–62.

Butcher, C., Goldin-Meadow, S., 2000. Gesture and the transition from one- to two-word speech: When hand and mouth come together. In: McNeill, D. (Eds.), Language and gesture. Cambridge University Press, New York, pp. 235–258.

Butterworth, B., Beattie, G., 1978. Gesture and silence as indicators of planning in speech. In: Campbell, R., Smith, G.T. (Eds.), Recent advances in the psychology of language: formal and experimental approaches. Plenum Press, New York, pp. 347–360.

Camaioni L., Perucchini, P., Bellagamba, F., Colonnesi, C., 2004. The role of declarative pointing in developing a theory of mind. Infancy 5, 291–308.

Capone, N. C., McGregor, K.K., 2004. Gesture development: A review for clinicians and researchers. J. Speech Lang. Hear. Res. 47, 173–186.

Cochet, H., Vauclair, J., 2010. Pointing gesture in young children: Hand preference and language development. Gesture 10, 129–149.

Colonnesi, C., Stams, G.J.J.M., Koster, I., Noom, M.J., 2010. The relationship between pointing gesture and language: A meta-analysis. Dev. Rev. 30, 352–366.

De Boysson-Bardies, B., Vihman, M.M., 1991. Adaptation to language: Evidence from babbling and first words in four languages. Language 67, 297–319.

De Ruiter, J.P., 1998. Gesture and speech production. Doctoral dissertation. Katholieke Universiteit, Nijmegen.

De Ruiter, J.P., 2000. The production of gesture and speech. In: McNeill, D. (Eds.), Language and Gesture. Cambridge University Press, Cambridge, pp. 284–311.

De Ruiter, J.P., 2006. Can gesticulation help aphasic people speak, or rather, communicate? Adv. Speech Lang. Pathol. 8 (2), 124–127.

Ejiri, K., Masataka, N., 2001. Co-occurrence of Preverbal Vocal behavior and Motor Action in Early Infancy. Dev. Sci. 4, 40–48.

Ekman, P., Friesen, W., 1969. The repertoire of nonverbal behavioural categories: Origins, usage and coding. Semiotica 1, 49–98.

Engstrand, O., Bonsdroff, L., 2004. Quantity and duration in early speech: preliminary observations on three Swedish children. In: Papers from the 17th Swedish Phonetics Conference, Stockholm, pp. 64–67.

Esteve-Gibert, N., Prieto, P., 2013. Intonational and gestural structures are temporally coordinated. J. Speech. Lang. Hear. Res. DOI 10.1044/1092-4388(2012/12-0049).

Esteve-Gibert, N., Prieto, P., 2012. Esteve-Prieto Catalan acquisition corpus. [http://prosodia.upf.edu/phon/ca/corpora/description/esteveprieto.html]

Feldman, R., Reznick, J.S., 1996. Maternal perception of infant intentionality at 4 and 8 months. Infant Behav. Dev. 19, 483–496.

Ferré, G., 2010. Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. In: Proceedings of Language Resources and Evaluation, Workshop on Multimodal Corpora, pp. 86–91.

Frota, S., Vigário, M., 2008. The intonation of one-word and first two-word utterances in European Portuguese. In: Proceedings of XIth International Congress for the Study of Child Language, Edinburgh.

Iverson, J.M., Fagan, M.K., 2004. Infant vocal-motor coordination: Precursor to the gesture-speech system? Child Dev. 75, 1053–1066.

Iverson, J.M., Goldin-Meadow, S., 2005. Gesture paves the way for language development. Psychol. Sci. 16, 367–371.

Iverson, J.M., Tencer, H.L., Lany, J., Goldin-Meadow, S., 2000. The relation between gesture and speech in congenitally blind and sighted language-learners. J. Nonverbal Behav. 24, 105–130.

Iverson, J.M., Thelen, E., 1999. Hand, mouth and brain: The dynamic emergence of speech and gesture. J. Conscious. Stud. 6, 19–40.

Kendon, A., 1980. Gesticulation and speech: Two aspects of the process of utterance. In: Key, M.R. (Eds.), The Relationship of Verbal and Nonverbal Communication. Mouton, The Hague, pp. 207–227.

Kendon, A., 2004. Gesture: Visible action as utterance. Cambridge University Press, Cambridge.

Kita, S., 2000. How representational gestures help speaking. In: McNeill, D. (Eds.), Language and gesture. Cambridge University Press, Cambridge, pp. 162–185.

Krahmer, E., Swerts, M., 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. J. Mem. Lang. 57 (3), 396–414.

Krauss, R.M., Chen, Y., Gottesman, R.F., 2000. Lexical gestures and lexical access: A process model. In: McNeill, D. (Eds.), Language and Gesture. Cambridge University Press, New York, pp. 261–283.

Lausberg, H., Sloetjes, H., 2009. Coding gestural behavior with the NEUROGES-ELAN system. Behav. Res. Methods Instrum. Comput. 41 (3), 841–849.

Leung, E.H.L., Rheingold, H.L., 1981. Development of pointing as a social gesture. Dev. Psychol. 17, 215–220.

Levelt, W.J., Richardson, G., La Heij, W., 1985. Pointing and voicing in deictic expressions. J. Mem. Lang. 24, 133–164.

Leonard, T., Cummins, F., 2010. The temporal relation between beat gestures and speech. Lang. Cogn. Process. 26, 1295–1309.

Liszkowski, U., 2007. Human twelve-months-olds point cooperatively to share interest with and helpfully provide information for a communicative partner. In: Liebal, K., Müller, C., Pieka, S. (Eds.), Gestural Communication in Nonhuman and Human Primates. John Benjamins, Amsterdam, pp. 124–140.

Loehr, D., 2007. Aspects of rhythm in gesture and speech. Gesture 7, 179–214.

Masataka, N., 2003. From index-finger extension to index-finger pointing: ontogenesis of pointing in preverbal infants. In: Kita, S. (Eds.), Pointing: Where Language, Culture, and Cognition Meet. Lawrence Erlbaum, Mahwah, pp. 69–84.

McNeill, D., 1992. Hand and Mind: What Gestures Reveal About Thought. The Chicago University Press, Chicago.

Melinger, A., Levelt, W., 2004. Gesture and the communicative intention of the speaker. Gesture 4, 119–141.

Nobe, S., 1996. Representational Gestures, Cognitive Rhythms, and Acoustic Aspects of Speech: A Network/Threshold Model of Gesture Production. Doctoral dissertation. University of Chicago, Chicago.

Oller, D.K., Wieman, L.A., Doyle, W.J., Ross, C., 1976. Infant babbling and speech. J. Child Lang. 3, 1–11.

Olson, S.L., Bates, J.E., Bayles, K., 1982. Predicting long-term developmental outcomes from maternal perceptions of infant and toddler behavior. Infant Behav. Dev. 12, 77–92.

Özçalişkan, S., Goldin-Meadow, S., 2005. Gesture is at the cutting edge of early language development. Cognition 96, 101–113.

Papaeliou, C.F., Trevarthen, C., 2006. Prelinguistic pitch patterns expressing ‘communication’ and ‘apprehension’. J. Child Lang. 33, 163–178.

Prieto, P., 2006. The relevance of metrical information in early prosodic word acquisition: a comparison of Catalan and Spanish. Lang. Speech 49 (2), 233–261.

Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R., Vanrell, M.M., 2013. Intonational phonology of Catalan and its dialectal varieties. In: Frota, S., Prieto, P. (Eds.), Intonational variation in Romance. Oxford University Press, Oxford.

Rochat, P., 2007. Intentional action arises from early reciprocal exchanges. Acta Psychol. 124, 8–25.

Rochet-Capellan, A., Laboissière, R., Galván, A., Schwartz, J.L., 2008. The speech focus position effect on jaw-finger coordination in a pointing task. J. Speech Lang. Hear. Res. 51 (6), 1507–1521.

Roustan, B., Dohen, M., 2010. Co-Production of Contrastive Prosodic Focus and Manual Gestures: Temporal Coordination and Effects on the Acoustic and Articulatory Correlates of Focus. In: Proceeding of the Speech Prosody, Chicago.

Rusiewicz, H.L., 2010. The Role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures. Doctoral dissertation. University of Pittsburgh, Pittsburgh.

So, W.C., Demir, O.E., Goldin-Meadow, S., 2010. When speech is ambiguous gesture steps in: Sensitivity to discourse-pragmatic principles in early childhood. Appl. Psycholinguist. 31, 209–224.

Sperry, L.A., Symons, F.J., 2003. Maternal Judgments of Intentionality in Young Children with Autism: The Effects of Diagnostic Information and Stereotyped Behavior. J. Autism Dev. Disord. 33 (3), 281–287.

Tomasello, M., Carpenter, M., Liszkowski, U., 2007. A new look at infant pointing. Child Dev. 78, 705–722.

Vanrell, M.M., Prieto, P., Astruc, L., Payne, E., Post, B., 2011. Adquisició dels patrons d'alineació i de camp tonal en català i espanyol. Anejo de Quaderns de Filologia, 71–88.

Vihman, M.M., DePaolis, R.A., Keren-Portnoy, T., 2009. Babbling and words: A Dynamic Systems perspective on phonological development. In: Bavin, E.L. (Eds.), The Cambridge Handbook of Child Language. Cambridge University Press, Cambridge, pp. 163–182.

Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., Miller, J., 1985. From babbling to speech. Language. 61 (2), 397–445.

West, B., Welch, K.B., Galecki, A.T., 2007. Linear mixed models: a practical guide using statistical software. Chapman & Hall/CRC, New York.