Nine-month-old infants are sensitive to the temporal alignment of prosodic and gesture prominences

Núria Esteve-Gibert,[a] Pilar Prieto,[b,a] & Ferran Pons[c,d]

[a]Universitat Pompeu Fabra, Department of Translation and Language Sciences, Spain

[b]Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain

[c]Departament de Psicologia Bàsica, Universitat de Barcelona, Spain

[d]Institute for Brain, Cognition and Behavior (IR3C), Barcelona, Spain

Abstract

This study investigated the sensitivity of 9-month-old infants to the alignment between prosodic and gesture prominences in pointing-speech combinations. Results revealed that the perception of prominence is multimodal and that infants are aware of the timing of gesture-speech combinations well before they can produce them.

Keywords: early temporal alignment; pointing; pointing-speech coordination; multimodal prominence; multimodal communication

Highlights:

- Nine-month-old infants perceive multimodal temporal alignment.
- Infants perceive prominence at both speech and gesture levels.
- The perception of temporally aligned gesture-speech combinations chronologically precedes their production.

Manuscript

When humans communicate we use multimodal cues (i.e., speech and gestures), which serve to transmit information more efficiently. Research over the past twenty years suggests that gesture and speech are tightly integrated in human communication semantically, pragmatically, and temporally (e.g., Birdwhistell, 1970; Kendon, 1980; Levinson, & Holler, 2014; McNeill, 1992). The semantic and pragmatic integration of gesture and speech refers to the fact that gestures that spontaneously co-occur with speech share the same meaning representation and the same communicative intention. The temporal integration (sometimes called 'phonological integration') entails that the most prominent part of co-speech gestures (i.e., the interval of the gesture stroke or the specific apex within the gesture stroke) co-occurs with the prosodically prominent part of speech (i.e., the accented syllable and the pitch peak within the accented syllable when available).

Regarding this last feature, temporal alignment in adults is evidenced by the fact that gesture and prosodic timing influence each other. First, speakers modify the acoustic realization of a word when it is accompanied by a visual beat (like a head nod, an eyebrow movement, or a manual beat gesture) in terms of duration and high vocalic formants. And, interestingly, listeners perceive a word as more prominent if it is accompanied by a visual beat (Krahmer & Swerts, 2007). Second, the prosodic structure of the target speech utterances influences the timing of associated gesture movements in the sense that prominent syllabic positions attract gesture prominences. In Catalan, for instance, the intonation peak of a focused pitch accent is typically associated with the most prominent syllable in a word; when a phrase boundary appears right after the stressed syllable, this pitch peak is retracted. And if this word is accompanied by a gesture, the gesture apex is also retracted, thus following the same pattern as intonation peaks (Esteve-Gibert & Prieto, 2013). This close temporal alignment has been observed in different co-speech gestures (manual deictic gestures, head movements, hand beat gestures, and articulatory gestures) using both naturalistic and experimental methods (De Ruiter, 2000; Leonard & Cummins, 2011; Levelt, Richardson, & La Heij, 1985; Loehr, 2012; McNeill, 1992; Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008; Roustan & Dohen, 2010; Rusiewicz, Shaiman, Iverson, & Szuminsky, 2013).

Speech perception studies on multimodal temporal alignment in infancy have mainly focused on infants' ability to detect the synchronization between lip movements and the corresponding audible speech (Lewkowicz, 2010; Pons & Lewkowicz, 2014). These studies have found that infants can detect an A-V desynchronization (with speech sounds and lip movements out of synchrony) in both isolated syllables (at 4 months of age) and in fluent speech (at 8 month of age). However, as far as we know, to date there are no studies that have investigated infants' early sensitivity to the integration of communicative gestures and speech. This information would help us understand how and when infants start being sensitive to the co-occurrence between gesture and speech prominences in order to be able to produce these combinations several months later.

Sensitivity to auditory prosodic speech prominence at the lexical level has been reported to appear very early on in language development. Infants can discriminate basic word stress patterns at the acoustic level from birth (Sansavini, Bertoncini, & Giovanelli, 1997). However, when more complex and variable stimuli are used, it is not until around 9 months of age that discrimination can be observed (Pons & Bosch, 2010; Skoruppa et al., 2009, 2013). Sensitivity to prominence marking has thus been explored at the auditory but not at the audiovisual level. In the present study we wanted to explore early sensitivity to audiovisual prominence marking by asking whether 9-month-old infants are sensitive to the temporal alignment between gesture prominence (i.e., the stroke) and the most prominent part of speech (i.e., prosodically accented syllables).

Crucially for our study, 9-month-old infants do not yet have the ability to combine communicative gestures with speech. It has been reported that rhythmic manual movements can be frequently coordinated with vocalizations in 6- to 11-month-old infants. However, these rhythmic movements are not yet communicative (Ejiri & Masataka, 2001). The first communicative gestures, pointing gestures, start being produced in isolation around 8-10 months of age (Bates, Camaioni, & Volterra, 1975) and it is not until around 15 months of age that infants combine most of their pointing gestures with speech (Butcher & Goldin-Meadow, 2000; Igualada, Bosch, & Prieto, under review; Murillo & Belinchón, 2012). In these gesture-speech combinations it seems that infants are already able to temporally align gesture and speech in an adult-like way, since 1) gestures start before the vocalizations associated with them, 2) the stroke onset coincides with the onset of the prominent syllable in speech, and 3) the gesture apex is produced before the end of the accented syllable (Esteve-Gibert & Prieto, 2014).

The current study is aimed at exploring whether 9-month-old infants are sensitive to the temporal alignment between gesture and speech (prosodic) prominences in co-speech pointing gestures. We predicted that infants would be sensitive to the alignment between prominences at this early age, long before they start producing temporally aligned pointing-speech combinations. If this is the case, three conclusions could be drawn from the study: (a) infants are sensitive to linguistic prominence not only at the auditory level but also at the gesture level, (b) the perception of multimodal temporal alignment comes long before its production, and (c) gesture and speech are already integrated in human communication – specifically, the temporal alignment of gesture and prosodic prominences– in early stages of development.

To verify our predictions, we tested twenty-four full-term 9-month-old Catalan-learning infants. The infants had an average age of 9.01 months (range: 256-287 days). Twelve additional infants were tested but not included in the final sample because of crying or fussiness (5 infants), failure to habituate (5), and experimental error (2). Participants were recruited at the maternity unit of the Hospital Sant Joan de Déu in Barcelona, Spain. Parental consent was obtained before running the experiment. The stimuli consisted of multimedia movies which were constructed using Premiere Pro CS5.5 (Adobe Corporation). The movies were video clips of a woman producing a pointing gesture accompanied by a disyllabic word produced in an infant-directed manner. The woman appeared facing sideways in the right-hand area of the screen, and then said a word while pointing to the left-hand part of the screen, at the same time covering her mouth with the hand not used for pointing to prevent infants from seeing her lip movements. Eighteen words in Catalan were used, half of them iambs and the other half trochees. All words were high-frequency words for talking to infants according to the MacArthur-Communicative Development Inventory (MCDI) (Fenson et al., 1994). There were two types of video clip: aligned clips, in which the gesture stroke (those video frames that capture maximum extension of the arm during the pointing gesture) coincided with the accented syllable of the pointing-accompanying word; and misaligned clips, in which the gesture stroke coincided with the unaccented syllable of the pointing-accompanying word. The misaligned clips were created using Premiere Pro CS5.5 by decoupling video and audio tracks and then displacing the video track backwards (in the case of iambs) or forwards (in the case of trochees) so that the gesture stroke coincided with the unaccented syllable.

Infants were tested in a dimly lit and sound-attenuated laboratory room, seated in a high chair facing a LG 50”TV screen at a distance of approximately 130 cm. The experiment was controlled by the experimenter from an adjacent room using Habit 2002 software (Cohen, Atkinson, & Chaput, 2000) running on a Power Mac G5. The infants’ looking behavior was video recorded for subsequent analysis. A habituation/test procedure was used to test for the detection of prosody-gesture alignment. The habituation phase consisted of the presentation of 15-second trials, each with three aligned video clips showing either all iambic or all trochaic stimuli (i.e., all words presented within a single habituation trial had the same stress type). The habituation criterion was set such that infant looking had to decline during a three-trial block to 60% of the total looking time observed during the longest block of three trials. When infants reached this criterion, the habituation phase ended and the test phase began. In the test phase four trials were presented, each consisting of four video clips. Two test trials contained aligned clips (one with iambic words and the other with trochees) and the other two test trials contained misaligned clips (again, one with iambic words and the other with trochees). These trials were presented in counterbalanced order across infants.

To determine whether infants were sensitive to the prosody-gesture misalignment we compared the duration of infants' looking time at each test trial. We submitted the data from the four test trials to a 2 × 2 × 4 mixed, repeated-measures ANOVA, with Stress Pattern and Alignment as within-subjects factors and test-trial order as the between-subjects factor. This analysis yielded only a significant main effect for Alignment ($F(1, 20) = 7.262$, $p = .014$, partial $\eta^2 = .266$), but not an interaction between these factors. These results indicate that infants detected the difference between the aligned and misaligned stimuli and that the detection of misaligned stimuli was not affected by the lexical stress pattern of the words (see Figure 1).

Figure 1

Following our predictions, two main contributions can be derived from our results. First, our results show that infants' early sensitivity to prominence is multimodal, because their ability to discriminate between word-initial and word-final prosodic prominence (Pons & Bosch, 2010; Skoruppa et al., 2009, 2013) can also be applied to their discrimination of pointing gesture prominences. Second, our results reveal that infants are aware of the adult-like timing of gesture-speech combinations several months before they actually produce these combinations, since infants do not start producing temporally aligned gesture and speech combinations until around 15 months of age (Butcher & Goldin-Meadow, 2000; Esteve-Gibert & Prieto, 2014). Thus, at 9 months of age infants not only know that prosodic prominence occurs at distinct positions within the word but can also notice the difference between stimuli in which the acoustic prominence coincides with gesture prominence and stimuli in which prominences do not co-occur.

All in all, our findings provide evidence for the early development of multimodal communication, which to date had been mainly studied in older children and adults. Future studies should explore whether the early sensitivity to the prosodic-gesture alignment seen in infants takes into account linguistic/semantic constraints or is based purely on the perception of systematic alignment patterns. Recent research with adults suggests that in natural discourse the timing of gesture-speech combinations can be influenced by the pragmatic coordination between the two modalities (Bergmann, Aksu, & Kopp, 2011; Esteve-Gibert, Pons, Bosch, & Prieto, 2014). In Esteve-Gibert et al. (2014), for instance, adults found that stimuli were acceptable when the gesture prominence occurred after the prosodic prominence, but not vice-versa, possibly because they interpreted each prominence as referring to a distinct speech act in the discourse. Although our results did not support the different role of stress pattern in prosodic-gesture combinations, it could be speculated that the infants' tendency to look longer at the misaligned iambic stimuli compared to the misaligned trochaic stimuli in our experiment could be a sign of a developing system in which the sensitivity to gesture-speech combinations is constrained by pragmatic, semantic, and linguistic factors. In addition, certain gestures such as negation gestures and emotion gestures marking ironic sentences do not seem to follow the prosodic timing constraints (González-Fuente, Escandell-Vidal, & Prieto, submitted; Harrison, 2010), and it also appears that different languages align gestures differently in speech depending on the semantics of the word involved (Alferink & Gullberg, 2014). Our study shows that infants as young as 9 months of age can detect misaligned stimuli, and we believe that this ability needs to be tested further in order to have a more complete picture of the early perception of multimodal temporal alignment.

Acknowledgments

We thank Helena Moliné for her assistance in collecting the data. We also thank the infants involved and their parent for their participation. This research was funded by the FFI2012-31995, 2014 SGR_925 and PSI2010-20294 research grants, and by the 2013 RECERCAIXA grant “Early precursors of language: an ICT guide for parents and caregivers”. A fuller report will be provided upon request.

References

Alferink, I., & Gullberg, M. (2014). French-Dutch bilinguals do not maintain obligatory semantic distinctions: Evidence from placement verbs. Bilingualism: Language and Cognition, 17, 22–37.

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. Merrill-Palmer Quarterly of Behavior and Development, 21(3), 205–226.

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, Germany.

Birdwhistell, R. L. (1970). Kinesics and Context: Essays on Body Motion Communication. Philadelphia: University of Pennsylvania Press.

Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech: when hand and mouth come together. In D. McNeill (Ed.), Language and Gesture (pp. 235–257). Chicago: Cambridge University Press.

Cohen, L. B., Atkinson, D. J., & Chaput, H. H. (2000). Habit 2000: A new program for testing infant perception and cognition. (Version 2.2.5c). Austin, Austin: University of Texas.

De Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and Gesture (pp. 284–311). Cambridge University Press. doi:10.1017/CBO9780511620850.018

Ejiri, K., & Masataka, N. (2001). Co-occurrence of preverbal vocal behavior and motor action in early infancy. Developmental Science, 4(1), 40-48.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements. Journal of Speech, Language, and Hearing Research, 56(850), 850–865.

Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. Speech Communication, 57, 301–316.

Esteve-Gibert, N., Pons, F., Bosch, L., & Prieto, P. (2014). Are gesture and prosodic prominences always coordinated? Evidence from perception and production. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), Proceedings of the Speech Prosody Conference (pp. 222–226). Dublin, Ireland.

Fenson, F., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in Early Communicative Development. Monographs of the Society for Research in Child Development, 59(5), 1–173.

González-Fuente, S., Escandell-Vidal, V., & Prieto, P. (submitted). Gestural codas lead to the interpretation of irony. Journal of Pragmatics.

Harrison, S. (2010). Evidence for node and scope of negation on coverbal gesture. Gesture, 10(1), 29–51.

Igualada, A., Bosch, L., & Prieto, P. (under review). Language development at 18 months is related to multimodal communicative strategies at 12 months. Infant Behavior and Development.

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

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

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. Language and Cognitive Processes, 26(10), 1457–1471.

Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and Voicing in Deictic Expressions. Journal of Memory and Language, 24, 133–164.

Levinson, S. C., Holler, J., (2014). The origin of human multi-modal communication. Philosophical Transactions of the Royal Society B, 369, 20130302.

Lewkowicz, D. J. (2010). Infant Perception of Audio-Visual Speech Synchrony. Developmental Psychology, 46(1), 66–77.

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology, 3, 71–89.

McNeill, D. (1992). Hand and mind: What gestures reveal about thought (p. 423). Chicago: University of Chicago Press.

Murillo, E., & Belinchón, M. (2012). Gestural-vocal coordination: Longitudinal changes and predictive value on early lexical development. Gesture, 12(1), 16–39.

Pons, F., & Bosch, L. (2010). Stress pattern preference in Spanish-learning infants: the role of syllable weight. Infancy, 15(3), 223–245.

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. Acta Psychologica, 149, 142–147.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J. (2008). The Speech Focus Position Effect on Jaw-Finger Coordination in a Pointing Task. Journal of Speech, Language, and Hearing Research, 51(6), 1507–1521.

Roustan, B., & Dohen, M. (2010). Gesture and Speech Coordination: The Influence of the Relationship Between Manual Gesture and Speech. Proceedings of INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association. Makuhari, Japan.

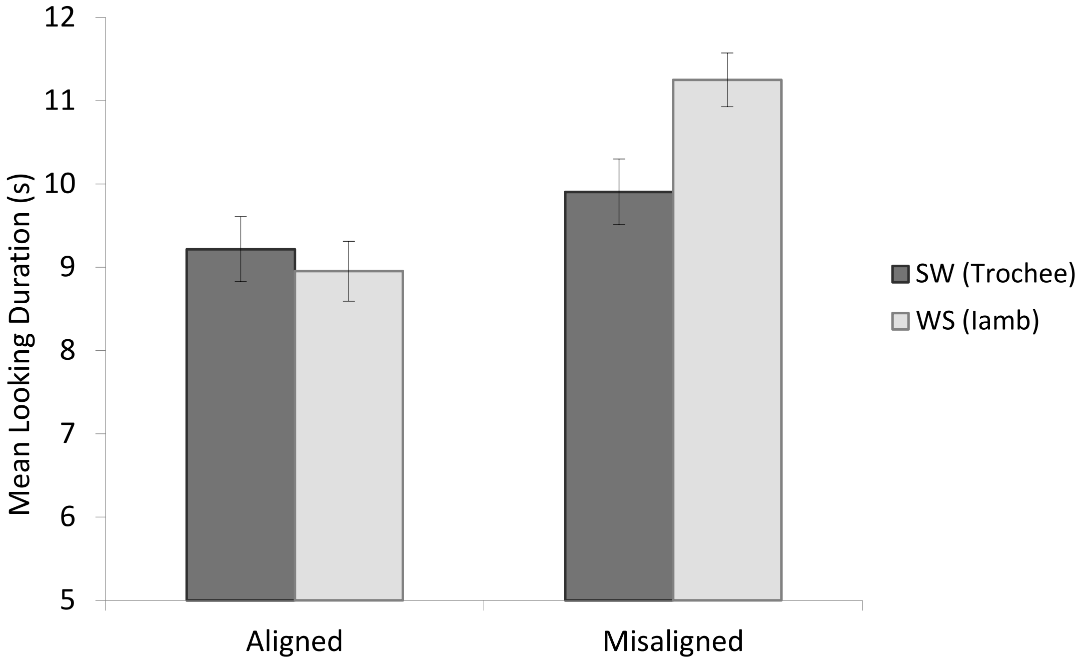
Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of Prosody and Position on the Timing of Deictic Gestures. Journal of Speech, Language, and Hearing Research, 56(2), 458–470.

Sansavini, A., Bertoncini, J., & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. Developmental Psychology, 33, 3–11.

Skoruppa, K., Pons, F., Bosch, L., Christophe, A., Cabrol, D., & Peperkamp, S. (2013). The Development of Word Stress Processing in French and Spanish Infants. Language Learning and Development, 9(1), 88–104.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., & Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. Developmental Science, 12(6), 914–919.

Figure 1. Mean looking times during test trials by 9-month-old Catalan-learning infants. Error bars represent standard error of the mean.


Mean Looking Duration (s)
Aligned
Misaligned
SW (Trochee)
WS (Iamb)