**Noname manuscript No.**
(will be inserted by the editor)

# How is information distributed across speech and gesture? A cognitive modeling approach

**Kirsten Bergmann · Sebastian Kahl · Stefan Kopp**

Received: date / Accepted: date

**Abstract** In naturally occurring speech and gesture, meaning occurs organized and distributed across the modalities in different ways. The underlying cognitive processes are largely unexplored. We propose a model based on activation spreading within dynamically shaped multimodal memories, in which coordination arises from the interplay of visuo-spatial and linguistically shaped representations under given cognitive resources. A sketch of this model is presented together with simulation results.

**Keywords** Speech · gesture · conceptualization · semantic coordination · cognitive modeling

## 1 Introduction

Gestures are an integral part of human communication and they are inseparably intertwined with speech [21]. The detailed nature of this connection, however, is still a matter of considerable debate. The data that underlie this debate have for the most part come from studies on the coordination of overt speech and gestures showing that the two modalities are coordinated in their *temporal* arrangement and in *meaning*, but with considerable variations. When occurring in temporal proximity, the two modalities express the same underlying idea, however, not necessarily identical aspects of it: Iconic gestures can be found to be *redundant* with the information encoded verbally (e.g., 'round cake' + gesture depicting a round shape), to *supplement* it (e.g., 'cake' + gesture depicting a round shape), or even to *complement* it (e.g., 'looks like this' + gesture depicting a round shape). These variations in meaning coordination–together with temporal synchrony–led to different hypotheses about how the two modalities encode aspects of meaning and what mutual influences between the two modalities could underlie this. However, a concrete picture of this and in particular of the underlying cognitive processes is still missing.

A couple of studies have investigated how the frequency and nature of gesturing, including its coordination with speech is influenced by cognitive factors. There is evidence that speakers indeed produce more gestures at moments of relatively high load on the conceptualization process for speaking [13, 22]. Moreover, supplementary gestures are more likely in cases of problems of speech production (e.g. disfluencies) or when the information conveyed is introduced into the dialogue (and thus conceptualized for the first time) [4]. Likewise, speakers are more likely to produce non-redundant gestures in face-tip-face dialogue as opposed to addressees who are not visible [2].

Chu et al. [8] provided data from an analysis of individual differences in gesture use demonstrating that poorer visual/spatial working memory is correlated with a higher frequency of representational gestures. However, despite this evidence, Hostetter and Alibali [10] report findings suggesting that speakers who have stronger visual-spatial skills than verbal skills produce higher rates of gestures than other speakers. A follow-up study

Kirsten Bergmann
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld
Tel.: +49-521-106-12143
E-mail: kirsten.bergmann@uni-bielefeld.de

Sebastian Kahl
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld
Tel.: +49-521-106-12947
E-mail: skahl@techfak.uni-bielefeld.de

Stefan Kopp
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld
Tel.: +49-521-106-12144
E-mail: skopp@techfak.uni-bielefeld.de

demonstrated that speakers with high spatial skills also produced a higher proportion of non-redundant gestures than other speakers, whereas verbal-dominant speakers tended to produce such gestures more in case of speech disfluencies [12]. Taken together this suggests that non-redundant gesture-speech combinations are the result of speakers having both strong spatial knowledge and weak verbal knowledge simultaneously, and avoiding the effort of transforming the one into the other.

In the literature, different models of speech and gesture production have been proposed. One major distinguishing feature is the point *where* in the production process cross-modal coordination can take place. The Growth Point Theory [21] assumes that gestures arise from idea units combining imagery and categorial content. Assuming that gestures are generated "prelinguistically", Krauss et al. [17] hold that the readily planned and executed gesture facilitates lexical retrieval through cross-modal priming. De Ruiter [24] proposed that speech-gesture coordination arises from a multimodal conceptualization process that selects the information to be expressed in each modality and assigns a perspective for the expression. Kita & Özyürek [14] agree that gesture and speech are two separate systems interacting during the conceptualization stage. Based on cross-linguistic evidence, their account holds that language shapes iconic gestures such that the content of a gesture is determined by bidirectional interactions between speech and gesture production processes at the level of conceptualization, i.e. the organization of meaning. Finally, Hostetter & Alibali [11] proposed the Gestures as Simulated Action framework that emphasizes how gestures may arise from an interplay of mental imagery, embodied simulations, and language production. According to this view, language production evokes enactive mental representations which give rise to motor activation.

Inspite of a consistent theoretical picture starting to emerge, many questions about the detailed mechanisms remain open. A promising approach to explicate and test hypotheses are cognitive models that allow for computational simulation. However, such modeling attempts for the production of speech and gestures are almost inexistent. Only Breslow et al. [7] proposed an integrated production model based on the cognitive architecture ACT-R [1]. This model, however, has difficulties to explain gestures that clearly complement or supplement verbally encoded meaning.

## 2 A Cognitive Model of Semantic Coordination

In recent and ongoing work we develop a model for multimodal conceptualization that accounts for the range of semantic coordination we see in real-life speech-gesture combinations. This account is embedded into a larger production model that comprises three stages: (1) conceptualization, where a *message generator* and an *image generator* work together to select and organize information to be encoded in speech and gesture, respectively; (2) formulation, where a *speech formulator* and a *gesture formulator* determine appropriate verbal and gestural forms for this; (3) *motor control* and *articulation* to finally execute the behaviors. Motor control, articulation, and formulation have been subject of earlier work [5]. In the following we provide a sketch of the model, details can be found in [15,3].

### 2.1 Multimodal Memory

The central component in our model is a multimodal memory which is accessible by modules of all processing stages. We assume that language production requires a preverbal message to be formulated in a symbolic-propositional representation that is linguistically shaped [19] (SPR, henceforth). During conceptualization the SPR, e.g., a function-argument structure denoting a spatial property of an object, needs to be extracted from visuo-spatial representations (VSR), i.e., the mental image of this object. We assume this process to involve the invocation and instantiation of memorized supramodal concepts (SMC, henceforth), e.g. the concept 'round' which links the corresponding visuo-spatial properties to a corresponding propositional denotation. Fig. 1 illustrates the overall relation of these tripartite multimodal memory structures.

To realize the VSR and part of the SMC, we employ a model of visuo-spatial imagery called *Imagistic Description Trees* (IDT) [25]. The IDT model unifies models from [20], [6], and [18] and was designed, based on empirical data, to cover the meaningful visuo-spatial features in shape-depicting iconic gestures. Each node in an IDT contains an imagistic description which holds a schema representing the shape of an object or object part. Important aspects include (1) a tree structure for shape decomposition, with abstracted object schemas as nodes, (2) extents in different dimensions as an approximation of shape, and (3) the possibility of dimensional information to be underspecified. The latter occurs, e.g., when the axes of an object schema cover less than the three dimensions of space or when an exact dimensional extent is left open but only a coarse relation between axes like "dominates" is given. This allows to represent the visuo-spatial properties of SMCs such as 'round', 'left-of' or 'longish'. Applying SMC to VSR is realized through graph unification and similarity matching between object schemas, yielding

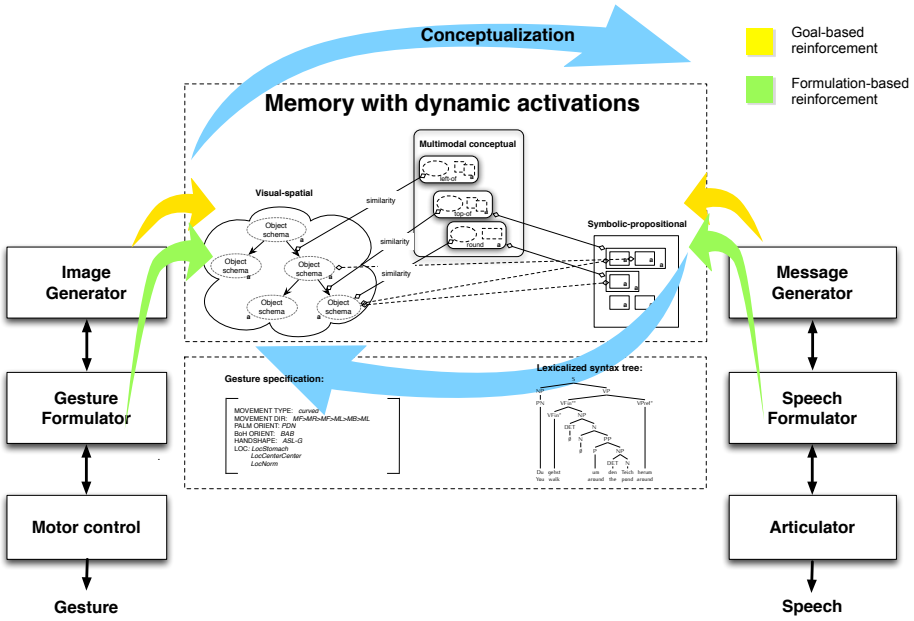


**Fig. 1** Overall production architecture.

similarity values that assess how well a certain SMC applies to a particular visuo-spatially represented entity (cf. Fig. 1). SPR are implemented straight forward as predicate-argument sentences.

### 2.2 Overall production process

Fig. 1 shows an outline of the overall production architecture. Conceptualization consists of cognitive processes that operate upon the abovementioned memory structures to create a, more or less coherent, multimodal message. These processes are constrained by principles of memory retrieval, which we assume can be modeled by principles of activation spreading [9]. As in cognitive architectures like ACT-R [1], activations float dynamically, spread across linked entities (in particular via SMCs), and decay over time. Activation of more complex SMCs are assumed to decay more slowly than activation in lower VSR or SPR.

Production starts with the *message generator* and *image generator* inducing local activations of modal entries, evoked by a communicative goal. VSRs that are sufficiently activated invoke matching SMCs, leading to an instantiation of SPRs representing the corresponding visuo-spatial knowledge in linguistically shaped ways. The *generators* independently select modal entries and pass them on to the *formulators*. As in ACT-R, highly activated features or concepts are more likely to be retrieved and thus to be encoded. Note that, as activation is dynamic, feature selection depends on the time of retrieval and thus available resources. The *message generator* has to map activated concepts in SPR onto grammatically determined categorical structures, anticipating what the *speech formulator* is able to process (cf. [19]). Importantly, interaction between *generators* and *formulators* in **each** modality can run top-down **and** bottom-up. For example, a proposition being encoded by the *speech formulator* results in reinforced activation of the concept in SPR, and thus increased activation of associated concepts in VSR.

In result, semantic coordination emerges from the local choices generators and formulators take, based on the activation dynamics in multimodally linked memory representations. Redundant speech and gesture result from focused activation of supramodally linked mental representations, whereas non-redundant speech and gesture arise when activations scatter over entries not connected via SMCs.

## 3 Results and outlook

To quantify our modeling results we ran simulation experiments in which we manipulated the available time (in terms of memory update cycles) before the model had to come up with a sentence and a gesture [15, 3]. We analyzed the resulting multimodal utterances with respect to semantic coordination: Supplementary (i.e., non-redundant) gestures were dominant in those

runs with stricter temporal limitations, while redundant ones become more likely when time available is increased. The model, thus, offers a natural account for the empirical finding that non-redundant gestures are more likely when conceptualization load is high, based on the assumption that memory-based cross-modal coordination consumes resources (memory, time), and is reduced or compromised when such resources are limited.

To enable a direct evaluation of our simulation results in comparison with empirical data, we currently conduct experiments to set up a reference data corpus. In this study, participants are engaged in a dyadic description task and we manipulate the preparation time available for utterance planning. The verbal output will subsequently be analyzed with respect to semantic coordination of speech and gestures based on a semantic feature coding approach as already applied in [4].

In ongoing work we extend the model to also account for *complementary* speech-gesture ensembles in which deictic expressions in speech refer to their co-speech gesture as in "the window looks like this". To this end, we advance and refine the feedback signals provided by the behavior generators to allow for the fine-grained coordination as it is necessary for the production of this kind of utterances. With this extension the model will allow to further investigate predictions as postulated in the lexical retrieval hypothesis [16, 23, 17]. Although that model was set up on the basis of empirical data, it was subject to much criticism based on psycholinguistic experiments and data. Data from detailed simulation experiments based on our cognitive model can provide further arguments in this debate.

## References


1. Anderson, J., Bothell, D., Byrne, M., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review **111**(4), 1036–1060 (2004)
2. Bavelas, J., Kenwood, C., Johnson, T., Philips, B.: An experimental study of when and how speakers use gestures to communicate. Gesture **2**(1), 1–17 (2002)
3. Bergmann, K., Kahl, S., Kopp, S.: Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In: R. Aylett, B. Krenn, C. Pelachaud, H. Shimodaira (eds.) Proceedings of the 13th International Conference on Intelligent Virtual Agents, pp. 203–216. Springer, Berlin/Heidelberg, Germany (2013)
4. Bergmann, K., Kopp, S.: Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In: Proceedings of SemDial2006, pp. 90–97 (2006)
5. Bergmann, K., Kopp, S.: GNetIc—Using Bayesian decision networks for iconic gesture generation. In: Proceedings of IVA 2009, pp. 76–89. Springer, Berlin/Heidelberg (2009)
6. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review **94**, 115–147 (1987)
7. Breslow, L., Harrison, A., Trafton, J.: Linguistic spatial gestures. In: Proceedings of Cognitive Modeling 2010, pp. 13–18 (2010)
8. Chu, M., Meyer, A.S., Foulkes, L., Kita, S.: Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. Journal of Experimental Psychology: General **143**(2), 694–709 (2013)
9. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review **82**(6), 407–428 (1975)
10. Hostetter, A., Alibali, M.: Raise your hand if you're spatial—relations between verbal and spatial skills and gesture production. Gesture **7**, 73–95 (2007)
11. Hostetter, A., Alibali, M.: Visible embodiment: Gestures as simulated action. Psychonomic Bulletin and Review **15/3**, 495–514 (2008)
12. Hostetter, A., Alibali, M.: Cognitive skills and gesture-speech redundancy. Gesture **11**(1), 40–60 (2011)
13. Kita, S., Davies, T.S.: Competing conceptual representations trigger co-speech representational gestures. Language and Cognitive Processes **24**(5), 761–775 (2009)
14. Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language **48**, 16–32 (2003)
15. Kopp, S., Bergmann, K., Kahl, S.: A spreading-activation model of the semantic coordination of speech and gesture. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci 2013), pp. 823–828. Cognitive Science Society, Austin, TX (2013)
16. Krauss, R., Chen, Y., Chawla, P.: Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Advances in Experimental Social Psychology **28**, 389–450 (1996)
17. Krauss, R., Chen, Y., Gottesman, R.: Lexical gestures and lexical access: A process model. In: D. McNeill (ed.) Language and gesture, pp. 261–283. Cambridge University Press, Cambridge, UK (2000)
18. Lang, E.: The semantics of dimensional designation of spatial objects. In: M. Bierwisch, E. Lang (eds.) Dimensional adjectives: Grammatical structure and conceptual interpretation, pp. 263–417. Springer, Berlin (1989)
19. Levelt, W.J.M.: Speaking: From intention to articulation. MIT Press (1989)
20. Marr, D., Nishihara, H.: Representation and recognition of the spatial organization of three-dimensional shapes. In: Proceedings of the Royal Socienty of London, vol. 200, pp. 269–294 (1978)
21. McNeill, D., Duncan, S.: Growth points in thinking-for-speaking. In: Language and gesture, pp. 141–161. Cambridge University Press, Cambridge, UK (2000)
22. Melinger, A., Kita, S.: Conceptualisation load triggers gesture production. Language and Cognitive Processes **22**(4), 473–500 (2007)
23. Rauscher, F., Krauss, R., Chen, Y.: Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science **7**, 226–231 (1996)
24. de Ruiter, J.: The production of gesture and speech. In: D. McNeill (ed.) Language and gesture, pp. 284–311. Cambridge University Press, Cambridge, UK (2000)
25. Sowa, T., Kopp, S.: A cognitive model for the representation and processing of shape-related gestures. In: Proc. European Cognitive Science Conference (2003)