Interaction between Speech and Gesture: Strategies for Pointing to Distant Objects

Thies Pfeiffer

A.I. Group, Faculty of Technology, Bielefeld University
Universitätsstr. 25, 33615 Bielefeld, Germany
thies.pfeiffer@uni-bielefeld.de
http://www.techfak.uni-bielefeld.de/~tpfeiffe/

Originally presented at the GW 2011. Proceedings published by Springer in 2012.
The original publication is available at www.springerlink.com.

Abstract. Referring to objects using multimodal deictic expressions is an important form of communication. This work addresses the question on how pragmatic factors affect content distribution between the modalities speech and gesture. This is done by analyzing a study on deictic pointing gestures to objects under two conditions: with and without speech. The relevant pragmatic factor was the distance to the referent object. As one main result two strategies were identified which were used by participants to adapt their gestures to the condition. This knowledge can be used, e.g., to improve the naturalness of pointing gestures employed by embodied conversational agents.

Keywords: object deixis, pointing, multimodal expressions

1 Introduction

Utterance meaning is achieved through the *partnership* of speech and gesture – this is how Kendon [1] expressed 1980 the view on the interaction between speech and gesture. This view is adopted here. A model of this partnership describing the individual contributions of the modalities is, however, still under research. Findings from Levelt et al. [2] suggest that for the production of multimodal deictic expressions this partnership manifests primarily during the planning of the utterance, while motor control is mainly ballistic. During the planning phase the speaker has to come up with a strategy on whether to produce a pointing gesture and if so, how much effort to put into it: should she provide the general direction, draw attention to the target area or indicate the target object directly?

The present work is inter alia motivated by the proposal of Cassell et al. [3], who suggested creating predictive models for embodied conversational agents (ECAs) to test theories on speech and gesture interaction. The results presented here contribute to a data-driven model for multimodal deictic expressions. The leading question is, whether there are observable interactions between speech and gesture during pointing tasks. Knowledge about these interactions would, on the one hand, provide deeper insights into the internal processes of the production of multimodal expressions and, on the other hand, it would provide valuable information on how to improve the naturalness of multimodal deictic expressions of ECAs.

2 Related Work

The interaction between speech and gesture can be investigated on several levels, from the temporal organization at a lower level to interactions on the lexical, semantic and pragmatic levels (see, e.g., McNeill [4]).

Highly relevant for the production of multimodal expressions, for example, is the temporal relationship between speech and gesture. Levelt et al. [2] investigated this temporal interdependency of speech and gesture to compare two theoretical positions: the temporal structure could be either pre-planned before motor execution (ballistic view) or it could be a result of a recurrent feedback loop on the level of motor execution (interactive view). The results of their studies were in favor of the ballistic view.

On the lexical level, Morrel-Samuels and Krauss [5] found for speech-related gestures that the word familiarity of the lexical affiliate of a gesture did influence both the duration of the gesture and its onset relatively to the lexical affiliate. This finding also supports an interaction at least at the level of motor planning.

Bergmann and Kopp [6] speak of a semantic synchrony of speech and gesture as a continuum where on the one extreme gestures and speech can encode content redundantly and on the other extreme they can encode content complementary. Bergmann and Kopp's work thereby is primarily focused on iconic gestures and on semantic content.

A short side note: In reality, we sometimes also find an additional extreme, where the encoding of speech and gesture conveys conflicting information. For example, when talking about the right hand and displaying the left hand instead – a confusion found quite often. There are many causes that might lead to conflicting information. It could be an erroneous encoding or an intended misinformation. It might also be the result of contextual differences in the representations of the interlocutors, e.g. regarding the selected frame of reference or based on cultural differences.

In the following, however, we will concentrate on the interaction between speech and gesture on the continuum from redundant to complementary information. We are going to address the issue of multimodal deictic expressions and set our scope on the influence of pragmatic factors and observable effects.

Knowledge about the interaction of speech and gesture and especially about the distribution of content is valuable for the generation of multimodal expressions in embodied conversational agents [3].

Kranstedt and Wachsmuth [7] suggested an incremental algorithm for the distribution of content in multimodal deictic expressions, which is an improved version of an algorithm for unimodal referring expressions proposed by Dale and Reiter [8]. They identified the distance of the referent object to the speaker as an important factor, which in their algorithm determines whether one of two types of pointing is used in the multimodal referring expression: object pointing or region pointing. They, however, only specify functional differences of these two types of pointing, but do not provide a description of perceivable differences in the way the gestures are produced.

An alternative approach was taken by van der Sluis [9], who uses a graph-based algorithm to generate the multimodal referring expression. This graph-based approach explicitly represents the costs for individual decisions during the content distribution. For example, using an uncommon noun might be more costly than naming a basic color; pointing gestures to small objects might be more costly than such towards larger ones. Van der Sluis also describes the concept of two types of pointing gestures: precise and imprecise gestures, which seem to be equivalent to the object pointing and region pointing introduced by Kranstedt and Wachsmuth [7]. However, she also does not provide a description of the observable differences between the different types of pointing gestures.

3 Study

The research question driving this exploratory study was, whether there are any observable effects showing an interaction between speech and gesture during pointing tasks based on pragmatic factors, i.e., based on the situative context of the interlocutors. For this, data from a previous study on manual pointing to objects of varying distances [10] was analyzed. In this previous study, which focused on the precision of manual pointing gestures, the participants were either being allowed or disallowed to use speech. The pragmatic factor which was expected to influence the interaction between speech and gesture was the distance of the referent object to the speaker.

3.1 Study Design

In the study, pairs of participants were engaged in a face-to-face pointing game (see Figure 1 a). This game was realized in two conditions, in the unimodal condition the participants were only allowed to use gestures when pointing to objects (*without-speech* condition), in the multimodal condition, they were allowed speech and gestures (*with-speech* condition). Each pair played the two conditions of the game. Order was balanced between pairs.

In the pointing game, each participant was assigned one of two roles: the *description giver*'s task was to describe an object in such a way that the other participant in the role of the *object identifier* was able to correctly identify the intended object. Altogether, there were 32 objects within each condition and the order of demonstration was controlled. The feedback of the object identifier was restricted to a simple pointing gesture to the identified object, which was then classified by the description giver as being right or wrong. Regardless of the result, the game proceeded to the next object, thus reparations were not allowed on purpose. These restrictions were put in place as to force the description giver to be accurate enough for the demonstrations and not to rely on interactive reparations.

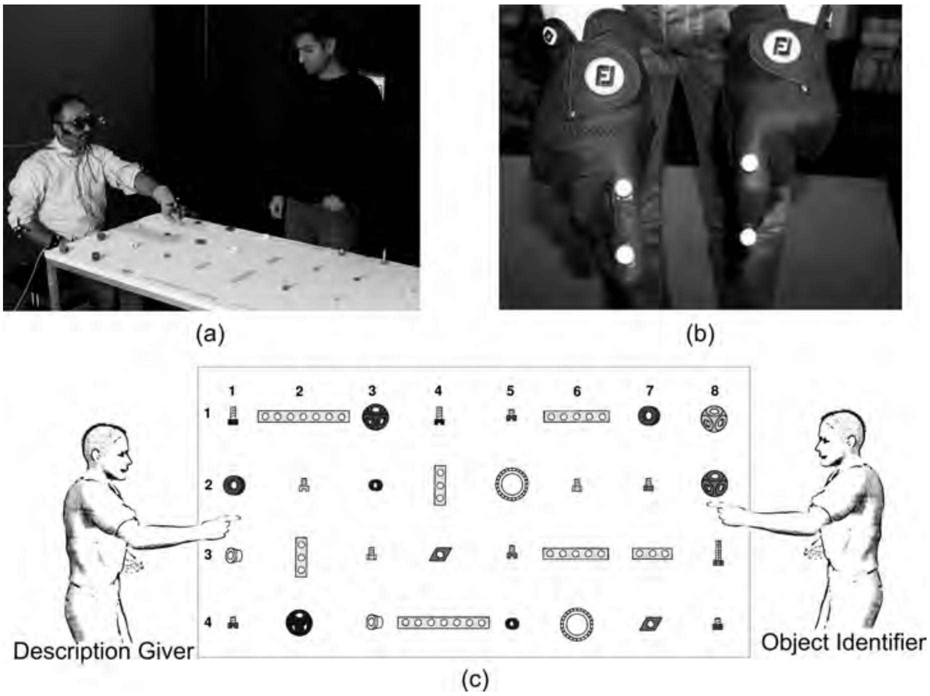


Fig. 1. The design of the study was based on a pointing game between two participants (a). Pointing gestures were recorded using optical tracking on the back of the hands (a) and on the index fingers (b). The target objects were arranged in a grid layout (c). The description giver was located to the left (in front of row 1), the object identifier to the right (behind row 8).

3.2 Domain

The objects of the target domain were parts of the wooden toy-kit Baufix. They had a distinctive shape and a basic planar coloring (one or two colors). In the preparation phase of the study, each participant was informed about the name of the objects to prevent misinterpretations.

The target domain consisted of 32 objects arranged in an uniform grid layout on top of a table (see Figure 1. c). The grid was arranged in 8 rows and 4 columns. The description giver was seated at the one side of the table just in front of row 1 and the object identifier was standing behind row 8. Columns 1 and 2 were to the left of the description giver, columns 3 and 4 to the right (vice versa for the object identifier).

3.3 Set-up

For the recording of the gestures, an optical tracking system was used (see the white spheres in Figure 1. a and b). The position and orientation of the description givers' index fingers, the back of the hands, the elbows and the head were recorded. During a pre-test, the hand gestures were tracked using a data-glove (see Figure 1. a). This was exchanged with an optical tracking focusing on the index fingers (see Figure 1. b) to improve the participant's acceptance of the tracking system. The object identifier was not tracked with the tracking system.

In addition to that, the scene was recorded by two video cameras (side view, top view). Speech was recorded using a headset microphone which was routed to one of the cameras. An additional microphone was placed above the scene and routed to the second camera.

The different recordings were synchronized using two displays showing the system time of the tracking system. Each display was visible from one video camera.

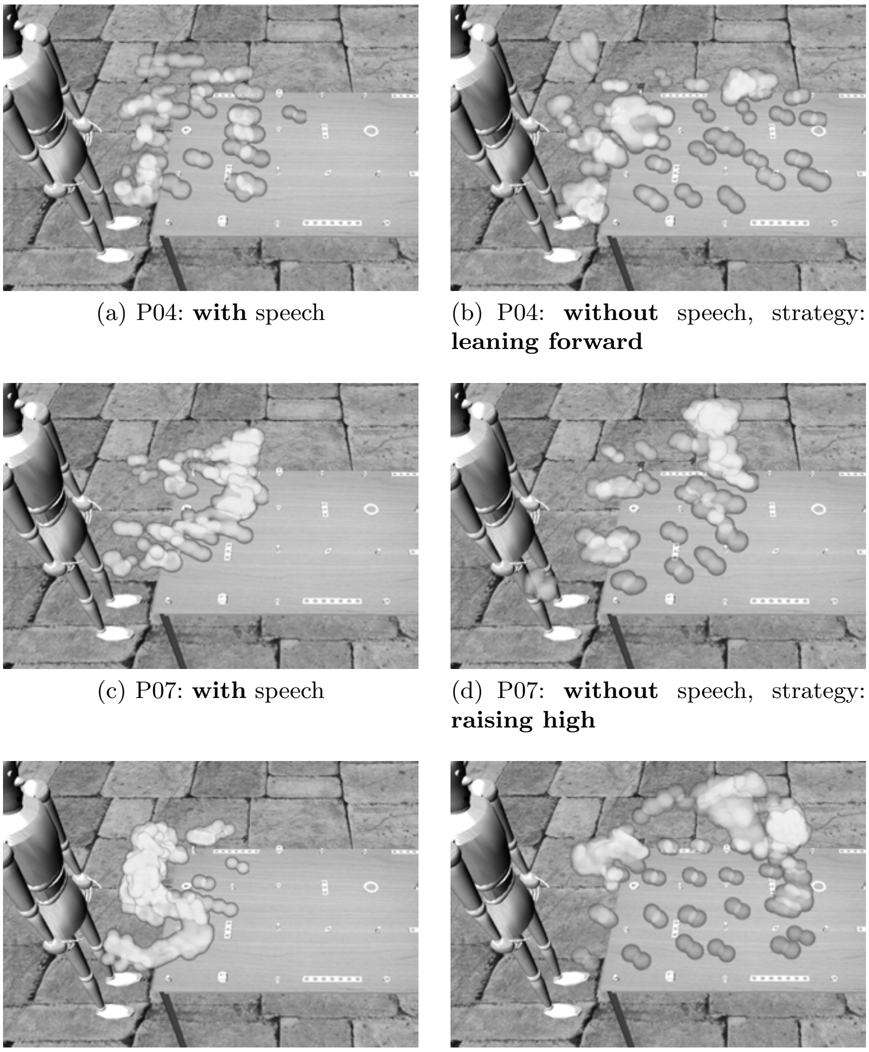
4 Results

In this study, for each of the 23 participants 32 multimodal deictic expressions were recorded for each of the two conditions. Altogether, 957 expressions with a manual deictic pointing gesture entered analysis.

4.1 Effects Found in Gesture

For the analysis of the participants' modulation of their manual pointing behavior under varying conditions, the positions of the tip of the elongated index finger during the strokes were visualized using Gesture Space Volumes (GSVs) [11]. In this case, GSVs encode the probability of positions as colors, similar to heatmaps, with red denoting high probability. Thus configured, GSVs provide a holistic perspective on the pointing behavior of an individual or a group by aggregating relevant multimodal deictic expressions in a single view while focusing on the end position during the stroke. Figure 2 and Figure 3 show examples of GSVs for different participants.

The GSVs for the different conditions (with speech and without speech) show, that the pointing behavior changes depending on the condition. Overall, the participants put more effort into their gestures when they were not allowed to speak. The participants thereby followed different strategies, which is also revealed by an analysis of the GSVs. Two main strategies emerge. When pointing to distant objects, 61% of the participants were **leaning forward** (see Figure 2.b) and 48% were **raising** their hands **high** above the table (see Figure 2.d). About 30% combined both strategies (see Figure 2.f). Other strategies involved increased dwelling (8%, see Figure 3.d) or frantic hand waving (4%, see Figure 3.b).



(a) P04: **with** speech

(b) P04: **without** speech, strategy: **leaning forward**

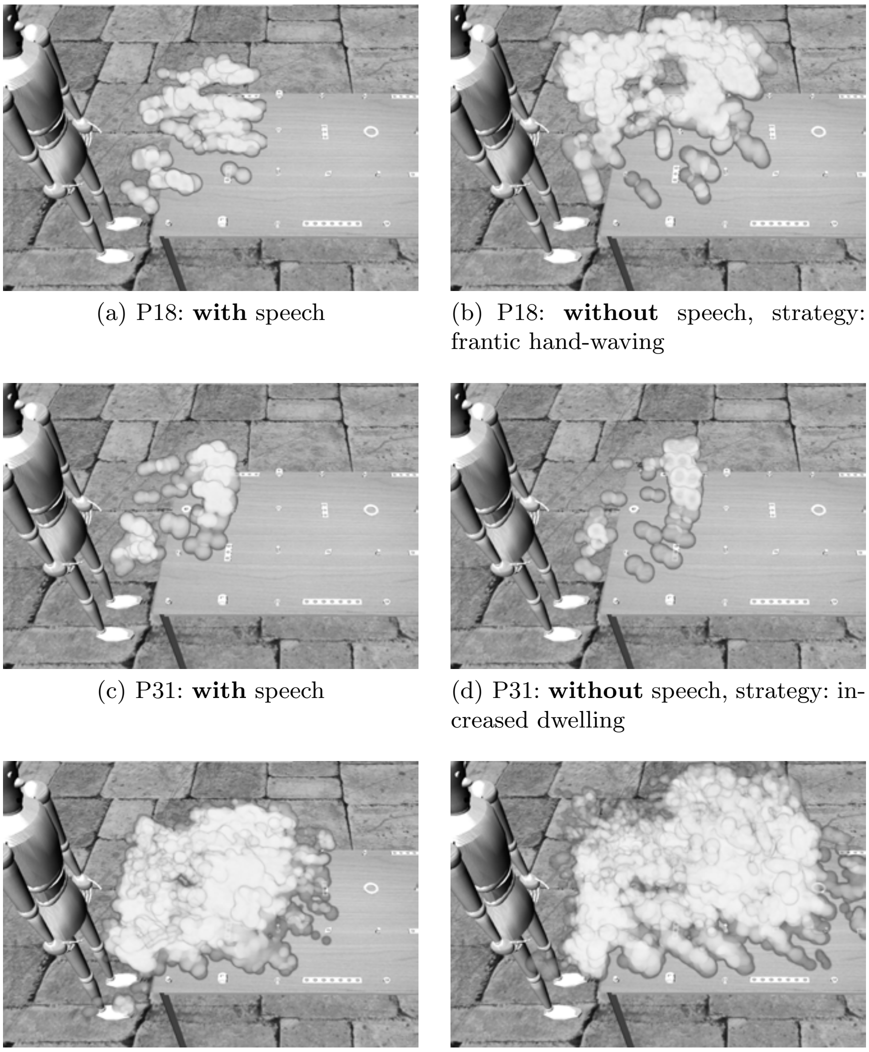
(c) P07: **with** speech

(d) P07: **without** speech, strategy: **raising high**

(e) P11: **with** speech

(f) P11: **without** speech, strategy: raising high and leaning forward

Fig. 2. Figures a)-f) show Gesture Space Volumes depicting the position of the tip of the index finger for each pointing act to the 32 objects in a setting. The wooden mannequin represents the position of the description giver.



(a) P18: **with** speech

(b) P18: **without** speech, strategy: frantic hand-waving

(c) P31: **with** speech

(d) P31: **without** speech, strategy: increased dwelling

(e) ALL: **with** speech

(f) ALL: **without** speech

Fig. 3. Less prominent strategies include *frantic hand-waving* (Figure b) and *increased dwelling* (Figure d). Figure f) depicts the Gesture Space Volume generated from the pointing acts of all participants taken together.

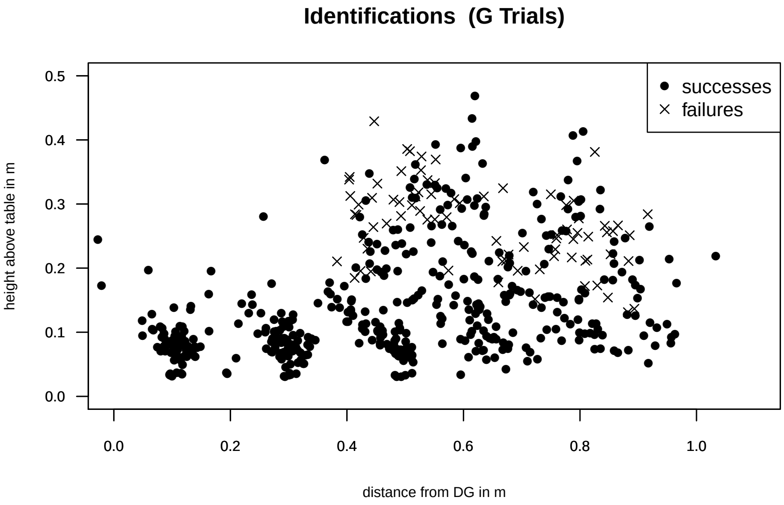


Fig. 4. The plot shows the positions of the index finger above the table over all participants for the trials without speech. The perspective is taken from the side of the table, the description giver is sitting to the left, the object identifier is standing to the right. Positions where the object identifier succeeded in determining the intended referent are marked with filled circles; those where the object identifier was not able to determine the intended referent are marked with a cross.

The multimodal deictic expressions were interpreted by a second participant opposite to the speaker. Those using **leaning forward** were correctly resolved, whereas the interpreters had problems with about one third of those using **raising high**. This is visualized in Figure 4, which depicts the success of an identification (by color) depending on the distance of the index finger of the description giver (increasing from left to right) and its height (increasing from bottom to top) during the stroke of the pointing gesture. Pointing gestures where the index finger hovered slightly above the surface of the table (indicated by the lower marks) are always identified successfully. Of the incidents where the index finger was higher than $0.1\,m$, however, more identifications failed, especially for cases when pointing to objects beyond $0.38\,m$. Hence the leaning-forward strategy with an index finger hovering slightly over the table was the most successful of the identified strategies.

4.2 Effects Found in Speech

When allowed to speak, participants increased the number of words used per multimodal deictic expression when referring to more distant objects (see Figure 5). This was evaluated by transcribing the recorded speech and counting the

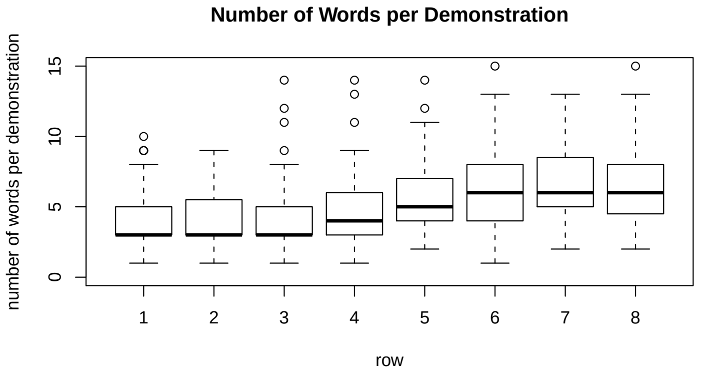


Fig. 5. Number of words used per multimodal expression as a function of the distance of the target object (given in rows from the speaker, see [10]).

numbers of words used in the verbal deictic expression associated with the manual pointing gesture. The most distant row 8 was a special case, as participants clearly showed an edge of the domain behavior, both in gesture and speech. At the same time, they put only little effort into the manual pointing gesture, i.e., they provided a general direction or indicated an area without moving the upper part of the body. For targets in the first three rows they could reach out effortlessly and nearly touch them with their finger tips. This is where the shortest verbal utterances were produced (e.g. "der rote Würfel"/"the red cube" (r3) vs. "die blaue Schraube hier außen in der Mitte"/"the blue bolt there in the middle of the outer side" (r8)).

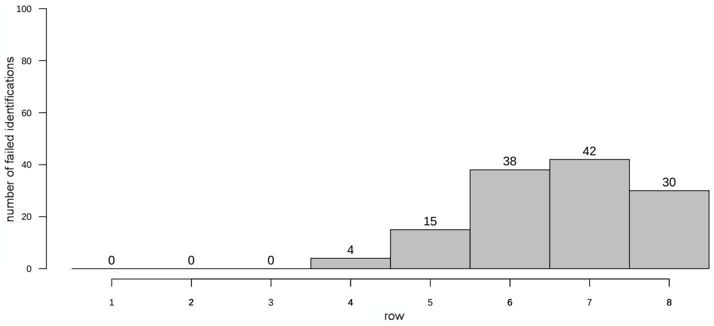


Fig. 6. The barplot shows the number of failed identifications per row for the trials without speech. Failed identifications increase from row 4 on.

That there is in fact an interaction between gesture and speech can be seen when comparing the increase in words with increasing distance depicted in Figure 5 with the increase in failures in identifying the targets of pointing gestures in the trials without speech depicted in Figure 6. Although taken from different conditions (with speech and without speech) and from different modalities, the general shape of both graphs is similar: a first increase between rows 3 and 4, a peak at row 7 and a slight degrees towards row 8.

Assuming that the identification success for the gestures produced in the with-speech condition are similar to the depicted identification successes for the without-speech condition, this suggests that the description giver is aware of the decreasing accuracy of the pointing gesture and tries to compensate for this by using more words in the referential expression.

5 Discussion

The analysis of the recorded data showed several interactions between speech and gesture. If participants had to rely on the manual gesture alone to refer to the objects, they put more effort into the gestures. This effort is guided by different strategies.

The dominant **leaning forward** strategy directly aims at reducing the distance between the relevant pointing device, the tip of the index finger, and the target object. The success of this strategy can be explained by an improved accuracy of the pointing gesture as perceived by the second participant. Given the spacing of the target objects, this led in all cases to a direct indication of the object.

With the **raising high** strategy, participants could have aimed at a better visibility of the pointing finger or they could have emphasized that the distant objects are behind others, exaggerating the effort to point over those closer objects. This strategy was, however, less successful than **leaning forward**. This can be attributed to an increased distance between the finger tip and the target object. In addition, raising the hand high above the target domain makes it difficult for the second participant to have both, pointing device and target object, within view simultaneously.

If, however, the participants were allowed to speak, they put less effort into their manual pointing gestures. Instead, if the distances to the target objects increased, they compensated this by using more words in their deictic expression.

The results suggest that the distance to the target object is a determining factor for both interactions. If the distance to the objects increases, the speaker has to put more effort into the multimodal deictic expression. The data suggests that there is a preference to put the effort into the verbal channel and only produce more expansive gestures if necessary. There is, however, a certain distance threshold. In the presented study this threshold is somewhere between row 3 and 4. The said interactions are found only for distances beyond this threshold. The threshold might be defined by the reach of the easily extended arm during a pointing gesture.

Bergmann and Kopp [6] found that gestures in path descriptions are more complementary and less redundant when verbal encoding is hard, e.g. when describing complex objects. Similar findings have been made by Bavelas et al. [12] for the description of pictures. Taking these findings related to semantic factors together with our own results for pragmatic factors, they suggest a conservative disposition of the speaker regarding the efforts invested in the multimodal expression. In the multimodal condition, the main contribution is encoded in the verbal channel and the gesture is redundant and/or rudimentary per default. If, however, the effort for the verbal encoding exceeds a certain threshold, more emphasis is laid on the gesture.

6 Conclusion

Starting with the question on the distribution of content between speech and gesture, individual strategies were identified which were employed by participants to modulate their pointing behavior when they were not allowed to speak. This was done using visual analysis of recorded tracking data based on *Gesture Space Volumes*. The results show that participants knew that the accuracy of pointing gestures decreases with the distance to the target object. If they were allowed to use speech and gesture, they compensated this loss of accuracy by increasing the number of words and thus avoided putting more effort into their gesture. If speech was not allowed, they had to increase the accuracy of their pointing gesture and did so by investing more effort into their gesture following two main strategies: leaning forward and raising high. Of these two, leaning forward was identified as being the most successful.

An important finding is that there is a threshold for the distance to the referent object, from which on the effort for producing the pointing gesture increases and that determines whether to increase the effort in the verbal channel instead. These insights can be used to inform the distribution of content when generating multimodal deictic expressions for ECAs. They should try to reduce the distance of their pointing gestures to the target object, but avoid, if possible, effortful movements, such as bending the upper part of the body. The latter, however, is necessary if speech is not allowed or not feasible or, which might be the case more often, the target object is difficult to discriminate using words.

In particular, the findings can be used to improve the algorithm of Kranstedt and Wachsmuth [7] by guiding the embodiment of their different types of pointing gestures in the following way: their object pointing can be mapped to the more expensive gestures, e.g. including the bending of the upper body, which reduce the distance to the referent object as much as possible, whereas their region pointing can be mapped to the more rudimentary, effortless gestures. For the algorithm of van der Sluis [9], the findings regarding the threshold from which on the effort of the pointing gesture seems to matter in gesture production can be used to fine tune the cost functions. Overall, by implementing the described strategies, the naturalness of the pointing behavior of embodied conversational agents should be increased.

References

1. Kendon, A.: Language and Gesture: Unity or Duality. In McNeill, D. (Ed): Language and gesture, pp. 47–63. Cambridge University Press (2000)
2. Levelt, W., Richardson, G., La Heij, W.: Pointing and Voicing in Deictic Expressions. Journal of Memory and Language, 24, Elsevier, 133–164 (1985)
3. Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., Badler, N., Pelachaud, C.: Modeling the Interaction between Speech and Gesture. In Ashwin Ram and Kurt Eiselt (Eds.): Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, pp. 153–158. Lawrence Erlbaum Associates (1994)
4. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press (1992)
5. Morrel-Samuels, P., Krauss, R.: Word Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 615–622 (1992)
6. Bergmann, K., Kopp, S.: Verbal or Visual? How Information is Distributed Across Speech and Gesture in Spatial Dialog. In Schlangen, D., Fernandez, R., (Eds.): Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pp. 90–97, Potsdam: Universitätsverlag (2006)
7. Kranstedt, A., Wachsmuth, I.: Incremental Generation of Multimodal Deixis Referring to Objects. Proceedings of the 10th European Workshop on Natural Language Generation (ENLG 2005), pp. 75–82, Aberdeen, UK (2005)
8. Dale, R., Reiter, E.: Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. Cognitive Science, 18, 233–263 (1995)
9. Van der Sluis, I.: Multimodal Reference - Studies in Automatic Generation of Multimodal Referring Expressions. PhD Thesis. BuG, Groningen (2005)
10. Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., Staudacher, M.: Measuring and Reconstructing Pointing in Visual Contexts. In Schlangen, D., Fernandez, R. (Eds.): Proceedings of the brandial 2006 - The 10th Workshop on the Semantics and Pragmatics of Dialogue, pp. 82–89. Universitätsverlag Potsdam, Potsdam (2006)
11. Pfeiffer, T.: Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces. PhD Thesis, Faculty of Technology, Bielefeld University. Aachen, Germany: Shaker Verlag (2011)
12. 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)