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/

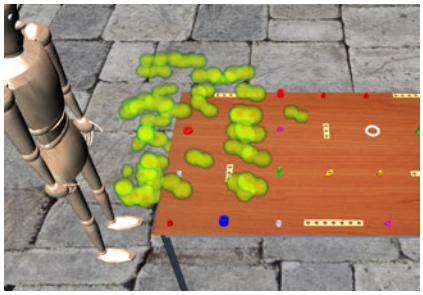
Abstract. Referring to objects using multimodal deictic expressions is an important form of communication. This work addresses the question on how content is distributed between the modalities speech and gesture by comparing deictic pointing gestures to objects with and without speech. As a result, two main strategies used by participants to adapt their gestures to the condition were identified. 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

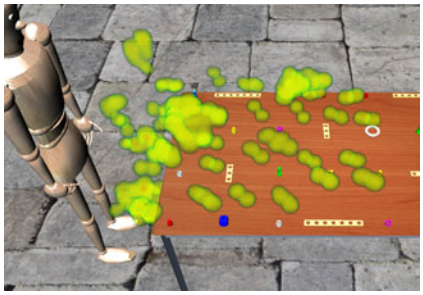
1 Introduction

Utterance meaning is achieved through the *partnership* of speech and gesture – this is how Kendon [2] 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. [4] suggest, that for the production of multimodal deictic expressions (MDEs), this partnership manifests primarily during the planning of the utterance, while motor control is mainly ballistic. It is during the planning phase, where 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?

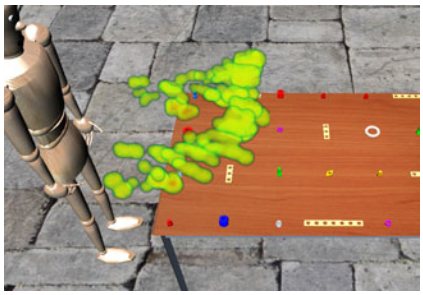
The present work is inter alia motivated by the proposal of Cassell et al. [1], who suggested creating predictive models for embodied conversational agents (ECAs) to test theories on speech and gesture interaction. The results presented here contribute to a data-driven model for MDEs. The leading question is how humans distribute content between speech and gesture to maintain a high discriminative power of MDEs under the constraint of an increasing distance to the target object.



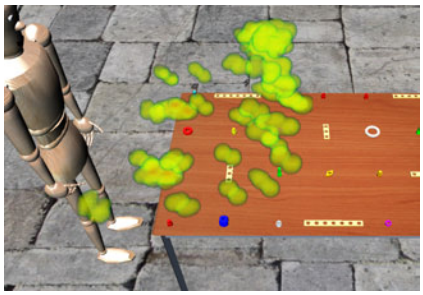
(a) P04: **with** speech



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



(c) P07: **with** speech



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

Fig. 1. Figures a)-d) 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 test person.

2 Study

The data used for this analysis was drawn from a study on manual pointing to objects of varying distances which was conducted with participants either being allowed or disallowed to use speech [3]. In this study, for each of the 23 participants 32 MDEs were recorded for each of the two conditions. Altogether, 957 expressions with a manual deictic pointing gesture entered analysis. Pointing gestures were recorded with high precision using an optical tracking system.

For the analysis of the participants' modulation of their 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) [5]. In this case, GSVs encode the probability of target 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 MDEs in a single view while focusing on the end position during the stroke. Fig. 1 shows examples from two participants.

3 Results

In the condition where participants were not allowed to speak, visual analysis of the GSVs revealed two main strategies. When pointing to distant objects, 61% of the participants were **leaning forward** (see Fig. 1.b) and 48% were **raising** their hands **high** above the table. About 30% combined both strategies. Other strategies involved increased dwelling (8%) or frantic hand waving (4%). The MDEs 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**.

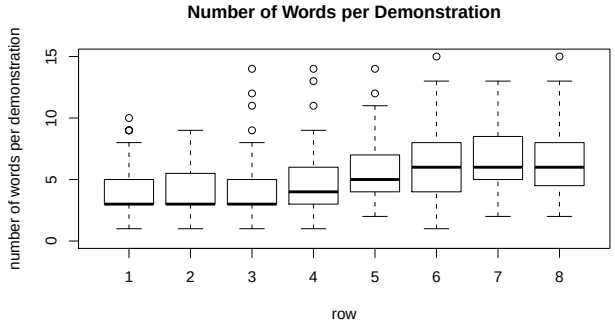


Fig. 2. Number of words used per multimodal expression as a function of the distance of the target object (given in rows from the speaker, see [3]).

When allowed to speak, participants increased the number of words used per MDE when referring to more distant objects (see Fig. 2). 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" (r3) vs. "die blaue Schraube hier außen in der Mitte" (r8)).

4 Discussion

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.

5 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 using two main strategies: leaning forward and raising high. Of these two, leaning forward was the most successful.

These insights can be used to inform the distribution of content when generating MDEs for ECAs. They should try to reduce the distance to the target object, but avoid 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.

References

1. 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).
2. Kendon, A.: Language and gesture: Unity or duality. In McNeill, D. (Ed): Language and gesture, pp. 47–63. Cambridge University Press (2000).
3. 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).
4. Levelt, W., Richardson, G., La Heij, W.: Pointing and voicing in deictic expressions. Journal of Memory and Language, 24, Elsevier, 133-164 (1985).
5. Pfeiffer, T.: Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces. Phd Thesis. Bielefeld University (2011, to appear).