# Deictic Gestures in Coaching Interactions

Iwan de Kok[1,2,3], Julian Hough[1,2], David Schlangen[1,2], Stefan Kopp[2,3]
[1]Dialogue Systems Group, [2]CITEC, [3]Social Cognitive Systems Group
Bielefeld University
idekok@techfak.uni-bielefeld.de

## ABSTRACT

In motor skill coaching interaction coaches use several techniques to improve the motor skill of the coachee. Through goal setting, explanations, instructions and feedback the coachee is motivated and guided to improve the motor skill. These verbal speech actions are often accompanied by iconic or deictic gestures and other nonverbal acts, such as demonstrations. We are building a virtual coach that is capable of the same behaviour.

In this paper we have taken a closer look at the form, type and timing of deictic gestures in our corpus of human-human coaching interactions. We show that a significant amount of the deictic gestures actually touch the referred object, that most of the gestures are complimentary (contrary to previous research) and often occur before the lexical affiliate.

## CCS Concepts

**•Computing methodologies → Discourse, dialogue and pragmatics;**

## Keywords

Multimodal Analysis; Deictic Gestures; Coaching

## 1. INTRODUCTION

The goal of our current line of research is to build a incremental multimodal coaching environment. In this system users go into a virtual reality environment and interact with a virtual coach teaching them how to do a motor skill, in our case a body-weight squat. Building a believable and effective virtual coach is a challenging task. Not only do we need to decide on what is currently wrong with the motor skill and what is best to give feedback on, but also in which way we need to address this problem. What is the best thing to say and do to make the user improve on the motor skill in focus?

To observe the behavior of human coaches in this situation we have recorded a corpus of human-human coaching interactions [7]. The coaches in this corpus single out an error and try to improve this error, before moving on to the next part of the motor skill. The instructions that are given are frequently accompanied with deictic gestures referring to the body parts in question.

In this paper we take a closer look at these deictic gestures in order to find out when, where and how to integrate them into our virtual coach. We are interested to learn more about what these deictic gestures look like, during what type of dialogue acts they occur and whether they complement what is being said or that they are redundant to the speech. Finally, the timing of the gestures in relation to the speech is also of interest.

The form and timing of deictic gestures has been researched before, but as far as we know not in the the domain for motor skill learning. The goal of this research is to either confirm the finding from other domains or learn the differences this domain brings.

In the following Section 2 we discuss the literature on deictic gestures. This is followed in Section 3 by a description of our corpus and the details of the annotation process. The results of our analyses are presented in Section 4 and we conclude with a discussion of the results in Section 5.

## 2. DEICTIC GESTURES

According to McNeill and Duncan [13] speech and gesture are synchronously produced to express the same underlying idea, but not necessarily to express the same aspects of it. In many cases the information is split over the verbal and nonverbal behavior and the full information becomes apparent only when the two modalities are combined– the modalities complement each other. In other cases the gesture or speech is redundant. The information that can be extracted from the gesture is the same as the information that can be extracted from the speech. Previous analyses of gesture have found either an even distribution of redundant and complimentary gestures [6, 18] or a tendency towards more redundant gestures [15, 1].

With respect to the temporal alignment of the gesture and the affiliated words of the gesture the research shows that in naturally occurring interaction the gesture either preceeds or synchronizes with the affiliated words [5, 16, 14, 3, 1]. There are several theories that explain the gap between gesture and speech. In the first theory the gap is explained by the difficulty in retrieving lexical items [5, 14, 11]. This theory advocates that gestures can assist in accessing the lexical expression via cross-modal priming. This requires that the gesture is already planned and in execution. The other



ACM ISBN 978-1-4503-2138-9.
DOI: http://dx.doi.org/10.1145/3011263.3011267

theory proposes that the gesture and speech are planned together [12, 8]. In this theory a common process distributes the information over the modalities in an early stage of the utterance production. The explanation that gestures are usually produced earlier is the that they lack the complexity of syntax and therefore need less production time. Secondly, according to this theory, image to gesture translation is an easier process than image to speech. Pointing at the object the gesture refers to is easier than finding the correct description for it. We are particularly interested in deixis for motor skill coaching, and to see whether these general findings on gesture apply to our domain. We are unaware of previous research on the topic in this domain.

# 3. PROCEDURE

In this section we present the corpus that is used for the analyses, the annotation made on this corpus and the process of forced alignment of the utterances to the speech to extract word timings.

## 3.1 Corpus

We invited 8 participants to interact with 2 different professional fitness coaches (4 participants per coach). We had one female (F) and one male coach (M). The goal of the interaction was to learn how to become competent at doing a body-weight squat. The coaches determined when the interaction ends, which was when they felt satisfied with the performance of the squat. The average length of the sessions is around 4:30 minutes. The participants had various different levels of expertise with squats ranging from novice to doing it on a monthly basis. None were professional athletes but all partook in recreational exercise. All sessions were in German and all participants were native German speakers.

## 3.2 Annotations

The dialogues were transcribed and annotated on dialogue acts, non-verbal gestures and actions using ELAN[1] [4] (see [9] for more details on the annotations). From these annotations we took the 114 *Deictic gesture* annotations and annotated them with the following information.

### 3.2.1 Target

Identification of the body part or object that was the target of the gesture. In case of the body parts a distinction is made between their own body parts (*self*) or those of the coachee (*other*).

### 3.2.2 Form

The deictic gestures are annotated distinguishing the following forms:

- **Pointing** - The coach points at the target with a stationary hand.
- **Area** - The coach highlights the target with a moving gesture.
- **Touch** - The coach touches the target.

The *touching* deictic gesture is very typical for our coaching dialogues, compared to other corpora looking at deictic gestures. It is especially useful when referring to the back of the coachee. The coachee can not see there him-/herself, so by touching the area the information can still be communicated.

Furthermore, the deictic gestures were annotated whether the gesture was performed with the left, right or both hands.

[1]http://tla.mpi.nl/tools/tla-tools/elan/


**Figure 1: ... und ich machs jetzt *[3]einmal vor. (... and I will show you once now.)**

### 3.2.3 Timing refinement

In the original annotations the onset of a gesture was the point in time where the hand(s) used in performing the gesture started moving. This includes the *preparation* phase where the hands move towards the starting location of the gesture. We needed more precision for our analysis, where we are interested in the time the coach is pointing at or highlighting the target of the gestures, the so-called *stroke hold* phase. Therefore, in a second pass over all gestures we identified the *stroke hold* phases of the gestures. In a few cases, where the *stroke* phase of the gesture was exaggerated and part of the intent of the gesture, the onset was identified as the start of the stroke phase. This occurred twice when the gesture was pointing towards a gaze target location and thus also used to simulate the gaze trajectory.

### 3.2.4 Type of Multimodality

The deictic gestures do not occur in a vacuum. These gestures are part of multimodal dialogue acts that transfer information from the coach to the coachee. There are several ways how this information can be split over the verbal and the nonverbal modality. It can be presented in only one of the modalities or both. Following Bergmann et al. [2], the deictic gestures in our corpus are categorized to reflect this division in the following ways:

- **Redundant** - The deictic gesture points at a target that was first mentioned by name in the speech accompanying the gesture.
- **Complimentary** - The deictic gestures points at a target that was first referred to in the speech accompanying the gesture, but not named (e.g. by using the word *'hier'* - *here*).
- **Supplementary** - The deictic gesture points at a target that was not mention nor referred to explicitly in the speech accompanying the gesture.

The supplementary deictic gesture is best illustrated by the extract in Figure 1.

During this quote the coach points at her own feet to direct the attention of the coachee toward the demonstration that follows. The gesture added the location of what will be

[3]* indicates the moment of the screenshot.

**Figure 2: ... hier nämlich eine neutrale Position in der * Wirbelsäule. (... here namely a neutral Position in the spinal column.)**

shown by the coach, which was not represented by the speech at all.

#### 3.2.5 *Lexical Affiliate*

For the *redundant* and *complimentary* deictic gestures there are words in the speech referring to the same body part or object as the target for the gesture, the lexical affiliates. For each of these gestures the first lexical affiliate is annotated. It was also the lexical affiliate that was used to distinguish between the *redundant* and *complementary* categories. Consider the following extract in Figure 2

In this case *hier* was chosen as the lexical affiliate and thus the deictic gesture pointing at the spinal column of the coachee was classified as a *Complimentary* gesture, even though the actual target was later mentioned by name ('Wirbelsäule' - spinal column).

### 3.3 Forced Alignment

The original transcripts of the corpus split the transcriptions into utterances; sentences or parts of sentences that formed a cohesive unit. However, for our analyses we are interested in the exact timing of the lexical affiliates. Therefore, we used forced alignment to get the individual word timings within an utterance. The original audio file was cut into individual utterances and along with the transcription given to WebMAUS [17, 10]. This gives us the estimated timing of the words at 10ms accuracy. All settings were left on the default values.

## 4. RESULTS

Table 1 presents an overview of the frequency of the different annotation labels for the analyzed deictic gestures.

As expected most of the deictic gestures target body parts (89%) of which 32% are their own body and 68% the body of the coachee. The other deictic gestures target a location in the room as a fixation point for the gaze during the squat movement. More than 40% of the deictic gestures touch the body part or object they are referring to.

Most deictic gestures are *Complimentary* (68), followed by *Redundant* (22). Only the female coach used *Supplementary* deictic gestures (6). Of the deictic gestures with lexical affiliate 75.6% were *Complimentary*, while the remaining 24.4% were *Redundant*.

Table 2 shows the frequencies of the different dialogue acts that co-occurred with the deictic gestures. It represents the dialogue acts present at the start of the deictic gesture. Most of the deictic gestures were during *Instructions* (47%), which is the most frequent dialogue act in our corpus (37%). Also, more often than expected is the co-occurrence with *Explanation* (11% vs 9%). *Feedback* co-occurs less often than

**Table 1: The counts of the different categories of deictic gestures**

| | Coach F | Coach M | Total |
|---|---|---|---|
| # Deictic Gestures | 69 | 45 | 114 |
| **Target** | | | |
| Own Body Part | 27 | 6 | 33 |
| Other's Body Part | 41 | 28 | 69 |
| Object | 1 | 11 | 12 |
| **Form** | | | |
| Pointing | 28 | 21 | 49 |
| Area | 13 | 8 | 21 |
| Touching | 29 | 17 | 46 |
| **Handedness** | | | |
| Left | 11 | 17 | 28 |
| Right | 20 | 23 | 43 |
| Both | 33 | 3 | 36 |
| **Multimodality** | | | |
| Redundant | 17 | 5 | 22 |
| Complimentary | 31 | 37 | 68 |
| Supplementary | 6 | 0 | 6 |
| Other | 11 | 5 | 16 |

**Table 2: The number of deictic gestures that started during each type of dialogue act.**

| | Coach F | Coach M | Total |
|---|---|---|---|
| **Start during** | | | |
| Instruction | 30 | 24 | 54 |
| Explanation | 11 | 2 | 13 |
| Feedback | 7 | 3 | 10 |
| Question | 4 | 0 | 4 |
| Commentary | 2 | 3 | 5 |
| SetGoal | 0 | 1 | 1 |
| None | 13 | 7 | 20 |

is to be expected (9% vs 14%).

For the 90 deictic gestures with a lexical affiliate the speech-gesture asynchrony was calculated: the differences between the onset of the gesture strokeholds and the onset of the lexical affiliate. Note, that in this analysis the strokehold timing was used - which comes later - as opposed to the stroke timing in the previous study [1]. Negative asynchrony values indicate that the lexical affiliate follows the gesture, positive asynchrony values indicate that the gesture follows the lexical affiliate.

Figure 3 illustrates the distribution of the speech-gesture asynchrony values found in our corpus. As can be seen for a majority of the instances the gestures preceded the lexical affiliate. Only 17 out of 90 (19%) of the deictic gestures started later than their lexical affiliate. The average speech-gesture asynchrony is -544 ms. This is quite a lot earlier than Bergmann et al. [1] report (-128 ms), even though they considered the *stroke* of the gesture the onset, which comes even earlier than the *strokehold* used here.

The outliers where the deictic gesture starts more than 2 seconds before the lexical affiliate are mostly due to hesitations in speech production in cases that is was easier to point at the object or direction. Take the example presented in Figure 4. In this case the coach is standing on the left side of and facing the towards the side of the coachee. The coach wants the person to take a small step away from the coach,

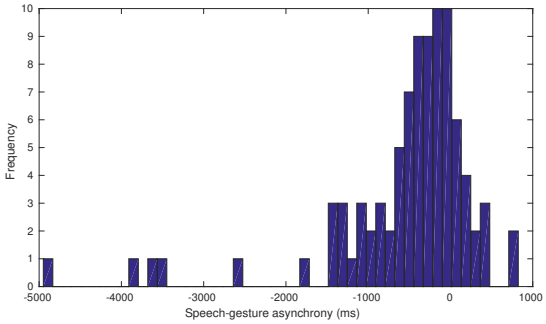


**Figure 3: Histogram illustrating the distribution of the speech-gesture asynchrony values.**


**Figure 4: ... ja geh mal einmal noch bitte ein ganz kleines Stuckchen weiter nach äh * (... okay please go a very tiny bit further to your uhm.) ... äh rechts * ist das von dir aus ja genau (... uhm right from your perspective yes exactly.)**

to the coachee's right. During this episode the coach makes three gestures pointing to the right of the coachee (second and third one are shown in the figure).

Here it is evident that pointing into desired direction is easier to produce, than finding the right formulation for the instruction. To determine whether the step the coachee needs to make is towards the left or right, the coach needs to mentally rotate - and during the third deictic gesture the coach does it physically as well - and translate herself to the position of the coachee and then make the desired step. Evidently, this led to a hesitation in the speech.

As final analysis we look a bit more closely at the temporal relations between the gesture and the lexical affiliate. For each gesture we looked whether the strokehold starts before, during or after the word and also whether the strokehold ends before, during or after the lexical affiliate. This results in six classifications in which a gesture-lexical-affiliate pair can fall. The results of this analysis is presented in Table 3.

In most cases the gestures covers the lexical affiliate 44 out of 90 (49%), meaning the gesture strokehold starts before the lexical affiliate and ends after. In most other cases (29 out of 90 (32%)) the gesture starts before the lexical affiliate, but ends earlier either before the lexical affiliate (13 out of 90 (14%)) or before the end of the lexical affiliate (16 out of 90 (18%)). The remainder start during the lexical affiliate and either end before the end of the lexical affiliate (1 out of 90 (1%)) or after the end of the lexical affiliate (10 out of 90) or start after the lexical affiliate all together (6 out of 90 (7%)).

**Table 3: Overview of the temporal relations between the gesture and the lexical affiliate. *Overlaps before* means the gesture starts before but ends during the lexical affiliate, while *overlaps after* means the gestures starts during but ends after the lexical affiliate.**

| | Coach F | Coach M | Total |
|---|---|---|---|
| **Gesture ... lexical affiliate** | | | |
| precedes | 4 | 9 | 13 |
| overlaps before | 5 | 11 | 16 |
| covers | 29 | 15 | 44 |
| fits within | 0 | 1 | 1 |
| overlaps after | 6 | 4 | 10 |
| follows | 4 | 2 | 6 |

## 5. DISCUSSION

In our corpus we have found more complementary gestures than redundant gestures. This is the opposite finding to the results of previous papers, where either an even distribution [6, 18] was found or more redundant gestures [15, 1]. Part of the difference may be explained that we only focused on deictic gestures and did not include iconic gestures.

Another difference of our study to the previous research is the domain. In our domain the coach gives precise instructions on the movement of the body. It can be very difficult to put what needs to be improved into words. The coach may know the technical terms of what needs to be improved, but these might be hard to understand by the novice coachee they interacted with in our corpus. We saw several attempts by the coaches to explain the same thing in varying ways, trying to get the point across. Sometimes the lexical form is not precise enough, especially regarding the spinal column. Here the coaches used touching deictic gestures to refer to precise location of the body. Demonstrations were also often used to show what cannot be said in words. Another difference with most of the other corpora is that the referred object is available in the environment of the interaction, contrary to a route giving task for instance. It is easy to point at or even touch, immediately removing any potential for confusion over what object is meant.

Our motivation for carrying out this analysis was the development of a virtual coaching character. Unfortunately, our virtual coach can not touch our user. Therefore, we are using other multimodal feedback strategies, such as highlighting body parts of the user in the virtual mirror.

However, regular deictic gestures without body contact are still useful for coaching a motor skill. We have seen that that these gestures often precede the lexical affiliate, in our domain even more than in previous research, even though we took a later point within the gestures as our onset. A likely explanation for this is that the coaches use it to guide the attention of the coachee to the problem area of the previous squat. In the decision making process, after each squat attempt by the coachee, the coach first makes a decision on which part of the squat to focus on next. By pointing at the area of the body where the error occurs, the coachee becomes immediately aware that the instructions and feedback that follow relate to that area of the squat. This is

particularly true for the 6 supplementary deictic gestures in our corpus. In all those, they were used to give a frame of reference for the instructions.

In the next phase we will integrate the deictic gestures in the behaviour of our virtual coach and evaluate their effectiveness. In particular, this analysis has given use insight in how to align the gestures with the lexical affiliates so we can replicate similar temporal relations in our system. Furthermore, we plan to investigate the features of the context which trigger the user of deixis in coaching interactions to gain more insight on when to include deixis and when to forego.

## 6. ACKNOWLEDGMENTS

We thank Cornelia Frank for help with the corpus collection, Angelika Maier for annotation and Gerdis Anderson, Michael Bartholdt and Oliver Eickmeyer for transcription. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

## 7. REFERENCES

[1] K. Bergmann, V. Aksu, and S. Kopp. The Relation of Speech and Gestures : Temporal Synchrony Follows Semantic Synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction*, number Ladewig, pages 1–6, 2011.

[2] K. Bergmann, S. Kahl, and S. Kopp. How is information distributed across speech and gesture ? A cognitive modeling approach. *Cognitive Processing, Special Issue: Proceedings of KogWis*, pages S84–S87, 2014.

[3] P. Bernardis and M. Gentilucci. Speech and gesture share the same communication system. *Neuropsychologia*, 44(2):178–190, 2006.

[4] H. Brugman and A. Russel. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068, 2004.

[5] B. Butterworth and G. Beattie. Gesture and silence as indicators of planning in speech. *Recent advances in the psychology of language: formal and experimental approaches*, pages 347–360, 1978.

[6] J. Cassell and S. Prevost. *Embodied Natural Language Generation: A Framework for Generating Speech and Gesture*. PhD thesis, 1997.

[7] I. de Kok, J. Hough, C. Frank, D. Schlangen, and S. Kopp. Dialogue structure of coaching sessions. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, pages 167–169, 2014.

[8] J. P. de Ruiter and D. McNeill. The production of gesture and speech. *Language and Gesture*, pages 284–311, 2000.

[9] J. Hough, I. de Kok, D. Schlangen, and S. Kopp. Timing and Grounding in Motor Skill Coaching Interaction : Consequences for the Information State. In *19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*, pages 86–94, 2015.

[10] T. Kisler, U. D. Reichel, F. Schiel, C. Draxler, and B. Jackl. BAS Speech Science Web Services âĂŞ an Update on Current Developments. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3880–3885, 2016.

[11] R. M. Krauss, Y. Chen, R. F. Gottesman, and D. McNeill. Lexical gestures and lexical access: a process model. *Language and Gesture*, pages 261–283, 2000.

[12] L. W. Levine. *The Unpredictable Past: Explorations in American Cultural History*. PhD thesis, 2011.

[13] D. Mcneill and S. D. Duncan. Growth Points in Thinking-for-Speaking. In *Language and gesture*, number 1987, chapter 7, pages 141–161. 1998.

[14] P. Morrel-Samuels and R. M. Krauss. Word Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3):615–622, 1992.

[15] K. J. Pine, N. Lufkin, E. Kirk, and D. Messer. A microgenetic analysis of the relationship between speech and gesture in children: Evidence for semantic and temporal asynchrony. *Language and Cognitive Processes*, 22(2):234–246, 2007.

[16] E. A. Schegloff. On some gestures' relation to talk. *Structures of social action*, pages 266–296, 1984.

[17] F. Schiel. Automatic phonetic transcription of nonprompted speech. In *Proceedings of ICPhS*, pages 607–610, 1999.

[18] H. Yan. *Paired Speech and Gesture Generation in Embodied Conversational Agents*. PhD thesis, 2000.