# Gestures and Self-Monitoring in Speech Production

MANDANA SEYFEDDINIPUR and SOTARO KITA
*Max Planck Institute for Psycholinguistics*

## 0. Introduction

Gesture and speech are semantically and temporally tightly coordinated. Gestures are pre-positioned temporally to the lexical affiliate with which they share semantic and/or pragmatic content (Kendon 1972, McNeill 1992, Schegloff 1984, Morrel-Samuels and Krauss 1992). The specific timing and semantic relation of speech and gesture has led to the view that gesture can serve as a "window" into mental processes underlying speech production (McNeill 1992).

In this paper we aim to gain insight into how speakers monitor their own speech by observing accompanying gestures. In natural speech, disfluencies are ubiquitous and they come in various forms. The flow of speech is interrupted by pauses and filled pauses (*uhm*), speakers search for specific words, they have trouble in the articulation of a word, and speakers also correct themselves. Self-corrections consist of different processes. Speech has to be monitored for appropriateness and for correctness. If an error is detected, the speech stream has to be interrupted, the repair has to be planned, and it then has to be finally executed.

The tight coordination between speech and gesture has led to the conclusion that (at least) at the conceptual level, speech and gesture production are closely interrelated. Furthermore, it has been argued that self-monitoring of speech is a conceptual level process (as opposed to a formulational level process, cf. Levelt 1989). Thus, it can be expected that gesture is sensitive to speech disfluency. By utilizing the specific timing relation of speech and gesture, we test two views regarding how speakers monitor their own speech.

We will investigate this issue on the basis of a corpus of disfluencies comprising descriptions of houses and apartments. The description of living spaces has proven to be a useful task for eliciting various kinds of gestures and speech disfluencies. The speaker has to transform three-dimensional space into the linear structure of speech. In addition, the speaker has to choose the appropriate words and constructions in order to convey the selected and linearized spatial information in a comprehensible way (Ullmer-Ehrich 1982). These difficulties result in a high number of disfluencies of different kinds and in a considerable use of gesture.

## 1. Monitoring Theories

Speakers monitor their own delivery constantly. They control their delivery such that what is going to be said is what they had intended. More specifically, they control for the appropriateness of selected words, and they check for errors (for details of foci of monitoring, see Levelt 1983, 1989). If an inappropriateness is detected, the speaker interrupts his speech stream and repairs the erroneous or inappropriate utterance. This whole process consists of four components: monitoring of speech, error detection, self-interruption, and self-correction. Various psycholinguistic theoretical accounts of speech monitoring and error detection have been proposed (for a review, see Postma 2000). Two of these accounts will be tested in this paper.

The INTERRUPTION-UPON-DETECTION HYPOTHESIS states that the speech stream is interrupted as soon as an error is detected. This is expressed in the Main Interruption Rule: "Stop the flow of speech immediately upon detecting trouble" (Levelt 1983, 1989; Nooteboom 1980). After the interruption of speech, the planning for reformulation takes place.

The rationale behind the Main Interruption Rule is that linguistic structures are ignored in interruption. Levelt's (1983) analysis showed that speakers interrupted their speech stream at any point in the delivery. They did not attend to any linguistic boundaries like syllables, words, or phrase boundaries. One exception is that speakers tended to complete non-erroneous words, i.e. neutral or merely inappropriate ones. This led to the refinement of the model such that the Main Interruption Rule only applies to cases of immediate detection of erroneous words.

The DELAYED-INTERRUPTION-FOR-PLANNING HYPOTHESIS suggests that even if an error is detected, the speaker does not interrupt his flow of speech immediately (Blackmer and Mitton 1991, Fox Tree and Clark 1997, Clark and Wasow 1998). Upon detection of an error the speaker will start the replanning and interrupt when the repair is ready to a certain degree or the speaker has run out of what can be uttered without further conceptual processing.

Blackmer and Mitton (1991) based their hypothesis on the analysis of the temporal characteristics of self-repairs in spontaneous speech. They observed that time intervals between the interruption point and resumption of speech were sometimes shorter than predicted by Levelt's Main Interruption Rule. According to the Main Interruption Rule, the replanning takes place only after the interruption. This implies that there has to be a time interval of some length before the resumption can take place. However, Blackmer and Mitton found instances where the suspension point was immediately followed by the correction, without any pause in between. Their results imply that the planning of the correction can take place while speaking is in progress and not only after suspension. Fox Tree and Clark (1997) came to a similar conclusion, but with a rather different type of evidence. They conducted a corpus study on the occurrence of the two pronunciation variants of the English article *the* (*thuh* with the reduced vowel schwa, and *thiy* with a non-reduced vowel). They found that 81% of the instances of *thiy* were followed by a suspension of speech. This suggests that speakers detected the

problem at some interval before suspending speech. By knowing in advance that they were going to suspend, the location of suspension (after *the*) and the type of suspension (the pronunciation of the variant *thiy*) is planned.

Taking the temporal and semantic interlocking of gesture and speech into account, the two theoretical approaches make different predictions concerning the gestural behavior. The Interruption-Upon-Detection Hypothesis predicts that any effect on gesture should be simultaneous with or follow the speech suspension. There should not be any effect on gesture before the actual speech suspension. This prediction is based on two assumptions: (i) when an error is detected, a stop signal is sent to both production modalities simultaneously (for an account of the suspension of speech and gesture production, see de Ruiter 2000), and (ii) it takes longer to suspend a gesture than speech because heavier mass has to be stopped in gesture.

In the case of the Delayed-Interruption-for-Planning Hypothesis, an effect on gesture can occur even before the moment of speech suspension due to the lag between error detection and speech suspension. When speakers have detected an error or have anticipated trouble, they start to plan how to resume right away and at the same time suspend the gestural movement. In the meantime, they go on speaking until the repair is ready up to a certain point or they have run out of formulated words. Consequently, gesture can stop before speech stops.

These predictions are tested by investigating the temporal relationship between different phases of self-repair and movement phases of gesture, which will be defined in the following sections.

## 3. Structural Components of Gesture and Speech Disfluencies

### 3.1. Disfluency Structure

A speech disfluency can be divided into different phases following Clark's (1996) disruption schema:

| | Suspension Point | Resumption Point | |
|---|---|---|---|
| **"On the right** | **\| uhm** | **\| on the left side...."** | |
| Original Delivery | Hiatus | Resumed Delivery | |

The first phase is the original delivery. The speaker monitors his internal speech for appropriateness and correctness (for a detailed description of foci of monitoring, see Levelt 1983). If an error is detected, the original delivery is disrupted. In the above example the original delivery is suspended at the word *right*. After the interruption a time interval (the hiatus) follows where speakers pause or utter filled pauses (e.g. *uhm*, *uh*) or so-called editing terms like *well*, *I think*, and *I mean*. The hiatus is seen as the phase where internal reformulation processes take place. The hiatus ends at the resumption point where the speaker resumes his delivery.

### 3.2. Gesture Structure

Gestures can be segmented into qualitatively different movement phases (Kendon 1972, 1980; McNeill 1992; Kita et al. 1998). The segmentation and identification of movement phases can be based purely on dynamic aspects of the hand/arm movement (as in Kita et al. 1998). In the preparatory movement phase the hands move from a resting position in order to prepare for the forcefully executed part, the stroke. The preparation phase can also be followed by a static phase, where the hands are held still in the initial position. This pre-stroke hold is then released by the stroke. The stroke phase is the semiotic and dynamic nucleus of the gesture. The stroke typically displays the meaning of the gesture. In the stroke the most force is exerted as compared to the neighboring phases. Also after this phase a static phase might follow, which is called the post-stroke hold. A gestural unit ends when the hands retract back into resting position, e.g. on the lap.

Preparation ⇒ Hold ⇒ **Stroke** ⇒ Hold ⇒ Retraction

Of the described gestural phases, only the stroke is obligatory. Note that in natural conversation one can observe a succession of strokes without the hands going into a hold or being retracted after each stroke.

## 4. Method

### 4.1. Stop Shifts and Start Shifts

In the analysis of the gestural movement pattern, we focused on the transition from one phase to another. Analogous to speech suspension and speech resumption, we distinguish two different types of phase shifts: a stop shift and a start shift. In a stop shift, an ongoing gestural unit/movement phase is suspended. In a start shift, a new dynamic gestural movement phase is initiated. These are described in more detail below.

**Stop shift**: an ongoing gestural movement is suspended or not completed.

- Shift of a dynamic phase into a static phase: an ongoing gestural movement phase (preparation/stroke) is suspended by going into a hold or by being retracted back into resting position.
- Shift of a dynamic phase into a new dynamic phase: a gesture gets suspended before being completed, e.g. a preparation phase is not followed by a stroke, for which the hands were preparing, but is followed by another preparation for the same or a different gesture.
- A dynamic phase is interrupted: a preparation or a stroke phase is prematurely truncated before the phase itself is terminated by a sudden abrupt halt or a sudden change in movement direction. In this case we classified the phase shift as a stop shift no matter what followed.

**Start shift**: a new gestural movement is started.

- Shift from a static phase into a dynamic phase: hands that are held still start a new preparation/stroke phase.
- A preparation phase is not followed by a stroke, but by a new preparation phase.
- An interrupted movement phase is followed by a new movement phase (preparation/stroke).

### 4.2. Data

The corpus consisted of six videotaped semi-natural conversations. Six native German speakers (four women, two men) were asked to describe houses and apartments they grew up in or had lived in for a longer period to a listener. Each session lasted 30-40 minutes. Nine minutes of the description from each speaker was transcribed. The speech data was coded for suspension points, hiatus length, and resumption points. The gestural movement phases were coded in terms of phase transitions. The temporal values were determined by a frame-by-frame microanalysis (1 frame = 40 ms). The six speakers produced a total of 582 disfluencies, of which 267 were overt repairs.[1] 191 overt repairs were accompanied by gestures, and 76 were not.

### 4.3. Analysis

We selected all utterances containing a repair that was accompanied by gestures (N=191). One speaker was excluded from the analysis because she did not provide sufficient data points. We analyzed the occurrences of stop shifts around suspension points and the occurrences of start shifts around resumption points. In order to ensure that the observations were independent from each other, we selected all repairs (i.e. the whole disfluency unit including suspension, hiatus, and resumption) that were at least two seconds apart from each other. We chose a time window of one second to each side of the suspension/resumption point and counted the number of start and stop shifts for every 160 ms slot within the window.

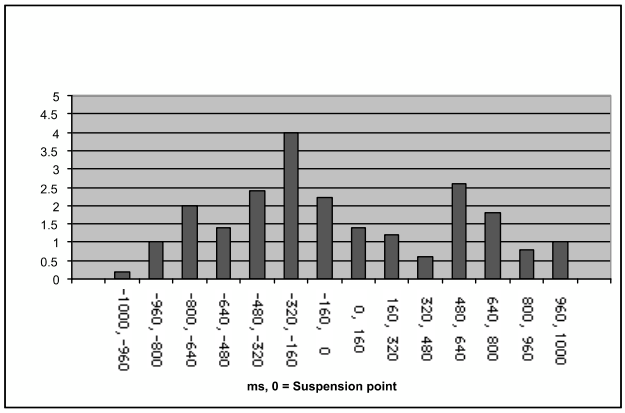
## 5. Results

### 5.1. Stop Shifts around the Suspension Point

Figure 1 presents the frequency of stop shifts around the speech suspension point (averaged over five speakers). The one-second window before and after the speech suspension point is divided into 160 ms intervals (0 ms = suspension point). Each bar shows the average frequency of stop shifts for a given time interval.

[1] Following Levelt (1983), we distinguish between overt and covert repairs. In a covert repair, the resumption is a continuation of the original delivery without any alternation (e.g. *the living room was, uhm, on the left side*). In contrast, an overt repair is an instance of a disfluency where in the resumption an element is altered (e.g. *the living room was, uhm, the dining room was on the left*).

**Figure 1.** Average frequency of stop shifts around suspension points

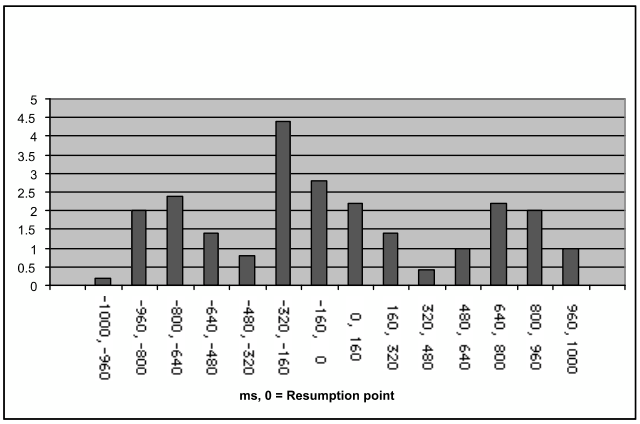



As is evident from Figure 1, gesture stops before speech stops. The most common time interval in which stop shifts occur is from -320 to -160 ms.

### 5.2. Start Shifts around the Resumption Point

Figure 2 presents the frequency of start shifts around the speech resumption point (averaged over five speakers). The one-second window before and after the speech resumption point is divided into 160 ms intervals (0 ms = resumption point). Each bar shows the average frequency of start shifts for a given time interval.

**Figure 2.** Average frequency of start shifts around resumption points



As is evident from Figure 2, gesture starts before speech starts. The most common time interval in which start shifts occur is from -320 to -160 ms.

## 6. Discussion

The above results show that gesture is highly sensitive to speech disfluencies. When speech is suspended and then resumed, gesture is also suspended and then resumed. Suspension and resumption in the two modalities are temporally coordinated in a systematic way. This suggests a highly interactive planning process that is involved in the production of both modalities.

Gesture is suspended prior to speech suspension. This suggests that gesture can be seen as an indicator of an upcoming interruption in speech. The gestural foreshadowing of speech suspension suggests that speakers are already aware that there is or will be trouble, but they do not interrupt speech right away. This is predicted by the Delayed-Interruption-for-Planning Hypothesis, according to which speakers continue speaking after error detection. They start planning for the resumption already before the speech suspension and disrupt their delivery when the repair is ready to a certain degree or they have run out of words that can be formulated without further conceptual planning. The above result also indicates that at least some utterances are interrupted in the way not predicted by the Interruption-Upon-Detection Hypothesis, according to which gesture should be interrupted simultaneously with or even after speech suspension.

However, these two hypotheses are not mutually exclusive. A speaker may interrupt his/her speech in different ways depending on various contextual factors. For example, in order to avoid losing the floor, one might delay suspension of speech. At the same time, in order not to mislead the interlocutor, one might suspend and repair the error as soon as possible. The speaker has to always evaluate advantages and disadvantages of speech suspension at a given moment. The timing of a speaker's interruption of his/her speech may be determined by a moment-by-moment balance among competing factors like comprehensibility and floor-keeping.

There is an emerging view in the literature that speech interruption is not a reflex-like reaction to error detection, but a choice the speaker makes based on, for example, the abovementioned factors (Blackmer and Mitton 1991, Fox Tree and Clark 1997, Clark and Wasow 1998). This study provides novel converging evidence for this idea by using speech-accompanying gesture as a window into the speaker's mind.

## References

Blackmer, Elizabeth R., and Janet L. Mitton. 1991. Theories of Monitoring and the Timing of Repairs in Spontaneous Speech. *Cognition* 39:173-194.

Clark, Herbert. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clark, Herbert, and Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. *Cognitive Psychology* 37:201-242.

de Ruiter, Jan-Peter. 2000. The Production of Gesture and Speech. In D. McNeill (ed.), *Language and Gesture*. Cambridge: Cambridge University Press, 284-311.

Fox Tree, Jean E., and Herbert Clark. 1997. Pronouncing "the" as "thee" to Signal Problems in Speaking. *Cognition* 62:151-167.

Kendon, Adam. 1972. Some Relationships between Body Motion and Speech. In A. Siegman and B. Pope (eds.), *Studies in Dyadic Communication*. New York: Pergamon Press, 177-210.

Kendon, Adam. 1980. Gesticulation and Speech: Two Aspects of the Process of Utterance. In M. R. Key (ed.), *The Relationship of Verbal and Nonverbal Communication*. The Hague: Mouton, 207-227.

Kita, Sotaro, Ingeborg van Gijn, and Harry van der Hulst. 1998. Movement Phases in Signs and Co-Speech Gestures, and Their Transcription by Human Coders. In I. Wachsmuth and M. Froehlich (eds.), *Proceedings of the International Gesture Workshop: Gesture and Sign Language in Human-Computer Interaction*. Bielefeld, Germany, 17-19 September 1997. Berlin: Springer, 23-35.

Levelt, Willem J. M. 1983. Monitoring and Self-Repair in Speech. *Cognition* 14:41-104.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

McNeill, David. 1992. *Hand and Mind*. Chicago: University of Chicago Press.

Morrel-Samuels, Palmer, and Robert M. Krauss. 1992. Word-Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:615-622.

Nooteboom, Sieb. 1980. Speaking and Unspeaking. Detection and Correction of Phonological and Lexical Errors in Spontaneous Speech. In V. Fromkin (ed.), *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand.* New York: Academic Press, 87-95.

Postma, Albert. 2000. Detections of Errors during Speech Production: A Review of Speech Monitoring Models. *Cognition* 77:97-131.

Shriberg, Liz. 1994. Preliminaries to a Theory of Speech Disfluencies. Ph.D. diss., University of California, Berkeley.

Max Planck Institute for Psycholinguistics
Wundtlaan 1
P.O. Box 310
6500 AH Nijmegen
The Netherlands

mandsey@mpi.nl
kita@mpi.nl