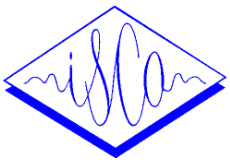

ITRW on Disfluency in
Spontaneous Speech (DiSS'01)
Edinburgh, Scotland, UK
August 29-31, 2001

Gesture as an Indicator of Early Error Detection in Self-Monitoring of Speech

Mandana Seyfeddinipur & Sotaro Kita

Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands
mandsey@mpi.nl

Abstract

There is a theoretical controversy regarding when the self-monitoring process interrupts the speech stream. One view holds that the speech stream is interrupted as soon as an error is detected. Another view holds that, even after an error is detected, the speaker does not interrupt immediately but continues speaking and at the same time plans the upcoming repair. We address this question by observing speech-accompanying gestures at the moment of speech disfluency. The results show that the concurrent gestural movements are typically stopped on average 240 ms before speech is stopped. In other words, the gesture suspension foreshadows the speech suspension. The gestural foreshadowing shows that the speaker must know early on that he is going to suspend speech. The gestural indication of an upcoming speech suspension suggests that the speaker does not interrupt speech at the very moment s/he detects an error. This result supports the hypothesis on speech monitoring stating that the speaker continues to talk after error detection and at the same time plans the upcoming repair.

1. Introduction

Gesture and speech are semantically and temporally tightly co-ordinated. For example, gestures are prepositioned temporally to the lexical affiliate, with which they share semantic and/or pragmatic content [1, 2, 3, 4]. The specific timing and semantic relation of speech and gesture has led to the view that gesture can serve as a ÒwindowÓ into mental processes underlying speech production [2]. In this paper, we aim to gain insight into how speakers monitor their own speech by observing accompanying gestures.

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, for example [5, 6, 7, 8]. Furthermore, it has been argued that self-monitoring of speech is a conceptual level process (as opposed to a formulational level process) [9]. Thus, it can be expected that gesture is sensitive to speech disfluency. By utilising 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 produced during on descriptions of houses and apartments. Living-space descriptions have proven to be a useful task in order to elicit various kinds of speech disfluencies and gestures. 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 linearised spatial information in a comprehensible way [10]. These difficulties result in a high number of disfluencies of different kinds and in a considerable use of gesture

2. Monitoring Theories

2.1. Two hypotheses about speech interruption

Speakers monitor their own delivery constantly. They control for 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 [9, 11]. If 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 psycho-linguistic theoretical accounts of error detection and self-interruption have been proposed (for a review, see [12]). 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: ÒStop the flow of speech immediately upon detecting troubleÓ [9, 11, 13]. 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. LeveltÕs [11] analysis showed that speaker 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 was that speakers tended to complete non-erroneous words, i.e. neutral or merely inappropriate ones. This led to the refinement of the model 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 [14, 15, 16]. 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 & Mitton [14] based their hypothesis on the analysis of the temporal characteristics of self-repairs in spontaneous speech. They observed time intervals between the interruption point and resumption of speech that were sometimes shorter than predicted by LeveltÕs [11] 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 & Mitton [14] 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 & Clark [15] 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 some interval before suspending speech. By knowing in advance that he is going to suspend, the location of suspension (after *the*), and the type of suspension (pronunciation of the variant *thiy*) is planned.

2.2. Predictions

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 following the speech suspension. There should not be any effect on gesture before the actual speech suspension. This prediction is based on two assumptions:
1. 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 [5])
2. It takes longer to suspend a gesture than speech because heavier mass has to be stopped in gesture.

According to 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 words that can be uttered without further conceptual processing. Consequently, gesture can stop before speech stops.

3. Phases in speech disfluency and gesture

The 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.

3.1. Disfluency structure

A speech disfluency can be divided into different phases following Clark [17] 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 [11]. 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 (*uhm, uh*) or so-called editing terms like *well, I think, I mean*. The hiatus is seen as the phase where internal reformulation processes take place [11]. 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 [1, 2, 18]. The segmentation and identification of movement phases can be based purely on dynamic aspects of the hand/arm movement [18]. 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, compared to the neighbouring 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

Schema: Gesture phases in a gesture unit

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. Data

The corpus consists of six videotaped semi-natural conversations. 6 native German speakers (4 female, 2 male) were asked to describe houses and apartments they grew up in or have 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 an overall of 582 disfluencies, of which 267 were overt repairs. 191 overt repairs were accompanied by gestures, and 76 were not.

4.2. Coding: 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.

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 a sudden abrupt halt or a sudden change in movement direction terminates the phase itself. 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).

4.3. Analysis

We selected all utterances containing a repair that was accompanied by gesture (191). One speaker was excluded from the analysis because she did not provide sufficient data points. We analysed 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 (the whole disfluency unit (suspension, hiatus, 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 points and counted the number of start and stop shifts for every 160 ms slot within the window.

5. Results

5.1. Stop shifts around 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 = suspension point). Each bar shows the average frequency of stop shifts for a given time interval.

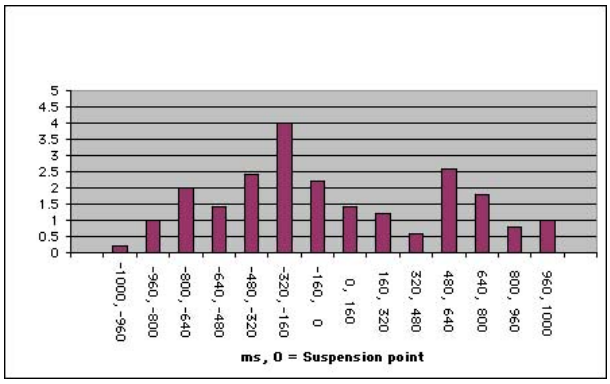

Figure 1 Average frequency of stop shifts around suspension points

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 (There is a secondary peak at around 400 ms to 800 ms after the suspension point The majority of these cases involve stopping of a new gesture that was generated after the suspension point. Thus, stop shifts in this peak are not directly related to the speech suspension).

5.2. Start shifts around resumption points

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 = resumption point). Each bar shows the average frequency of start shifts for a given time interval.

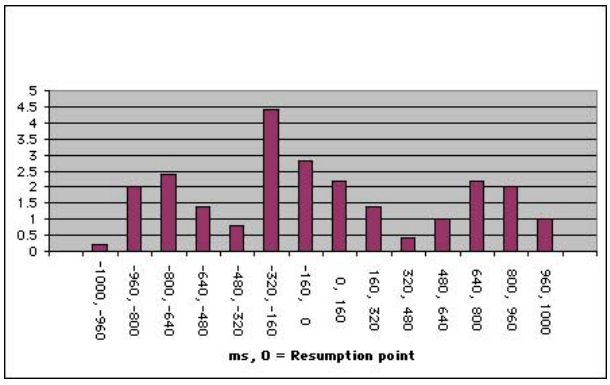

Figure. 2. Average frequency of start shifts around resumption points

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 (There are two secondary peaks —960 to —640 and from 640 to 960. The majority of gestures that are generated around these peaks are additional to the ones around the resumption point. Thus, start shifts in these peaks are not directly related to the resumption of speech).

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 the 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 the speaker is already aware that there is / will be trouble but he does 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 indicate 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. A moment-by-moment balance among competing factors like comprehensibility and floor keeping may determine the timing, at which a speaker interrupts his/her speech.

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, above mentioned factors [14, 15 16]. This study provides novel converging evidence for this idea by using speech-accompanying gesture as a window into the speakerÕs mind.

7. References

[1] A. Kendon, ÒSome relationships between body motion and speech,Ó in *Studies in dyadic communication,* A. Siegman and B. Pope, Eds. New York: Pergamon Press, 1972, pp. 177-210.

[2] D. McNeill, *Hand and mind*, Chicago: University of Chicago Press, 1992.

[3] E. Schegloff, Ò$n some gesturesÕ in relation to speech,Ó in *Structures of social action,* J.M. Atkinson and J. Heritage, Eds., Cambridge: Cambridge University Press, 1984, pp. 266-296.

[4] P. Morrel-Samuels and R. Krauss, ÒWord-familiarity predicts temporal asynchrony of hand gestures and speech,Ó *Jounal of Experimental Psychology: Learning Memory and Cognition,* vol. 18, pp. 615-622, 1992.

[5] J.P. de Ruiter, ÒThe production of gesture and speech,Ó in *Language and gesture,* D. McNeill Ed. Cambridge, Cambridge University Press, 2000, pp. 284-311.

[6] M.W. Alibali, S. Kita, and A.J. Young, ÒGesture and the process of speech production: we think, therefore we gesture,Ó *Language and Cognitive Processes*, vol. 15, pp. 593-613, 2000

[7] S. Kita, ÒHow representational gestures help speaking,Ó in *Language and gesture,* D. McNeill Ed. Cambridge, Cambridge University Press, 2000, pp. 162-185.

[8] U. Hadar and B. Butterworth, ÒIconic gestures, imagery and word retrieval in speech,Ó *Semiotica*, vol. 115, pp. 147-172, 1997.

[9] W.J.M. Levelt, *Speaking: From intention to articulation*, Cambridge Mass: MIT Press, 1989.

[10] V. Ullmer-Ehrich, ÒThe structure of living space descriptions,Ó in *Speech, place and action*, R.J. Jarvella and W. Klein Eds. Chichester: John Wiley, 1982, pp. 219-249.

[11] W.J.M. Levelt, ÒMonitoring and self-repair in speechÓ *Cognition,* vol. 14, 1983, pp. 41-104.

[12] A. Postma, Detection of errors during speech production: a review of speech monitoring models, *Cognition,* vol. 77, pp. 97-131, 2000.

[13] S. Nooteboom, ÒSpeaking and unspeaking. Detection and correction of phonological and lexical errors in spontaneous speech,Ó in *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, V. Fromkin, Ed. New York: Academic Press, pp.87-95, 1980.

[14] E.R. Blackmer and J.L. Mitton, ÒTheories of monitoring and the timing of repairs in spontaneous speech,Ó *Cognition*, 39, pp. 173-194, 1991.

[15] J.E. Fox Tree and H. Clark, ÒPronouncing ÒtheÓ as ÒtheeÓ to signal problems in speaking,Ó *Cognition*, vol. 62, pp. 151-167, 1997.

[16] H. Clark and T. Wasow, ÒRepeating words in spontaneous speech,Ó *Cognitive Psychology*, vol. 37, pp. 201-242, 1998

[17] H. Clark, *Using Language,* Cambridge: Cambridge University Press, 1996.

[18] S. Kita, I. van Gijn, and H. van der Hulst, ÒMovement phases in signs and co-speech gestures, and their transcription by human coders,Ó in *Gesture and sign language in human-computer interaction,* I. Wachsmuth and M. Froehlich Eds. Proceedings of the International Gesture Workshop, Bielefeld, Germany, September 17-19, 1997. Berlin: Springer, 1998, pp. 23-35.