Stellingen

behorende bij het proefschrift *Gesture and Speech Production.*

1. Het is in het algemeen niet mogelijk om de betekenis van spontane gebaren direct af te beelden op de begeleidende spraak. (dit proefschrift)

2. De *Growth Point* theorie van McNeill (1992) is circulair. Een Growth Point wordt gedetecteerd op basis van de synchronisatie van gebaar en spraak, en dient tegelijkertijd als verklaring van deze synchronisatie. (dit proefschrift)

3. Het maken van gebaren tijdens het spreken vergemakkelijkt het ophalen van informatie uit het visuele geheugen. (dit proefschrift)

4. Een belangrijke functie van het ontwikkelen van een informatieverwerkingsmodel is het in kaart brengen van de berekeningen die een rol spelen in het bestudeerde gedrag. (dit proefschrift)

5. Afgezien van de eis van berekenbaarheid is de enige beperking die informatieverwerkingmodellen opleggen aan de inhoud van een theorie dat het niet is toegestaan om een "homunculus" te poneren. (dit proefschrift)

6. De een z'n ruis is de ander z'n resultaat.

7. De in de informatica als *lazy* aangeduide evaluatiestrategie is mede zo inefficiënt omdat voortdurend berekend moet worden wat er allemaal niet berekend hoeft te worden.

8. Uit onderzoek in de vergelijkende taalkunde blijkt duidelijk dat in warme landen interessantere talen worden gesproken dan in koude landen.

GESTURE AND SPEECH PRODUCTION

MPI SERIES IN PSYCHOLINGUISTICS

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing
   Miranda van Turennout

2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography
   Niels O. Schiller

3. Lexical access in the production of ellipsis and pronouns
   Bernadette M. Schmitt

4. The open–/closed–class distinction in spoken–word recognition
   Alette Haveman

5. The acquisition of phonetic categories in young infants: A self–organising artificial neural network approach
   Kay Behnke

6. Gesture and speech production
   Jan-Peter de Ruiter

ISBN: 90–76203–05–9

Design: Linda van den Akker, Inge Doehring

Cover illustration: Inge Doehring

Printed and bound by: Ponsen & Looijen bv, Wageningen

GESTURE AND SPEECH PRODUCTION

een wetenschappelijke proeve
op het gebied van de Sociale Wetenschappen

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op donderdag 19 februari 1998
des namiddags om 1:30 uur precies

door

Jan-Peter de Ruiter

geboren op 13 oktober 1964 te Leiden

Promotor:	Prof. dr. W.J.M. Levelt
Manuscriptcommissie:	Prof. dr. H.H.J. Kolk Prof. dr. C.H.M. Gussenhoven Dr. S. Kita

The research reported in this thesis was supported by a grant from the Max–Planck–Gesellschaft zur Förderung der Wissenschaften, München, Germany.

To my father, who taught me to think.

ACKNOWLEDGEMENTS

A large number of people have contributed to this dissertation. First of all, I would like to thank Pim Levelt and Steve Levinson for giving me the opportunity to write my dissertation on such a fascinating subject. Their support and guidance during this project was indispensable.

I am indebted to the members of the Gesture Project for their support during the research and writing of my dissertation. Sotaro Kita, Pim Levelt, Steve Levinson, Eric Pederson, Gunter Senft, and David Wilkins listened patiently to my discourses on millisecond measurements and gave many thoughful comments and constructive criticisms. The same holds for the members of the gesture project in a more global sense of the word: I wish to thank Martha Alibali, Evelyn McClave, Sue Duncan, Robert Krauss, Adam Kendon, Rachel Mayberry , David McNeill, and Aslı Özyürek for the inspiring discussions we had about gesture and speech.

Furthermore, I wish to thank the members of the Cognitive Anthropology group for the fact that they "adopted" me into their building during the reconstruction of the Institute.

Since this was a technically very challenging project, I am indebted to the members of the Technical Group at MPI for their support. Specifically I wish to thank Dik van den Born, Rainer Dirksmeyer, Christa Hausmann-Jamin, Gerd Klaas, and Peter Wittenburg for providing great hardware and software support. I also want to thank Inge Doehring for the beautiful illustrations she made, both for the cover design and for the study in chapter 3.

I would like to take this opportunity to mention some friends without whom this project would have been significantly less fun. Theo Vosse entertained me with his literary email letters and taught me how computers really work, Marc Fleischeurs and Kay Behnke introduced me to Emacs, Linux and LaTeX, the coffee breaks with Dik van den Born prevented me from going crazy while I was analyzing video tapes, David Wilkins made editing the Annual Report together almost a pleasant job, and Eric Pederson (who also gave me the cute title for chapter 3) shared

many ideas, lunches, experiences and a hotel room in Albuquerque with me.

Finally, I want to thank my wife Suze not only for her extensive proof-reading of my dissertation, but also for her enduring support throughout these hectic years.

CONTENTS

INTRODUCTION

CHAPTER 1

The famous bestseller "Body Language" (Fast, 1971), the book that popularized the science of nonverbal behavior, has had a severe impact on the social life of a gesture researcher. When the subject of gesture is mentioned, people will often claim to know all about it, because they have read "Body Language". Although the book sometimes mentions gestures, they are clearly not its subject. The speech-related hand gestures that are the subject of this thesis differ from general nonverbal behavior in interesting ways.

The main difference between gesture and other nonverbal behavior such as gaze, posture or facial expression is that gestures are, in fact, quite "verbal". First, they are produced only during speaking[1]. Furthermore, the content (or "meaning") of gestures is usually directly related to the content of the concurrent speech. It is as if the voice and the hands are telling the same story in different ways (McNeill, 1985).

Deaf people rely on a system of gestures as their primary means of communication. However, the sign language of the deaf is a real language, in the sense that it is structured like a natural language: it has a syntax, morphology, and lexicalized semantics. Although subject to certain conventions and lexicalized forms (such as the thumbs-up gesture), speech-related gesture, as a semiotic system, does not have the prop-

[1] People also gesture when they want to speak but can't, for instance in noisy environments. This specific use of gesture is discussed in chapter 2.

erties of a real language, such as American Sign Language or Dutch. Nevertheless, the meanings we convey in gesture are mostly directly related to the accompanying speech. Gesturing is intricately related to speaking, and it is this relation that is the subject of this thesis.

1.1 Classification of Gestures

Gestures come in all sorts, and have accordingly been classified in many different ways. Rimé and Schiaratura (1991) give an extensive overview of a number of classification systems in the literature. The classification system adopted here is by McNeill (1992), for it is widely used, and it incorporates important distinctions that were also made in the influential classifications by Efron (1941) and Ekman and Friesen (1969). In McNeill's scheme, the following major categories can be distinguished.

Deictic gestures are gestures that refer to locations or directions. Most deictic gestures are pointing gestures, but other body movements (e.g. a head nod) can also be deictic gestures, if the intention of the speaker is to single out a location or direction. McNeill distinguishes between *abstract* and *concrete* deictic gestures. Concrete deictic gestures are gestures that single out an object or direction that is in the physical world. Abstract deictic gestures create or refer to discourse markers in the "gesture space" in front of the speaker's body. A frequently occurring abstract deictic gesture is when someone says "On the one hand ..." [open hand moves to the right] " and on the other ..." [same open hand moves to the left].

Iconic gestures are gestures whose shape resembles their referent in the speech. An iconic gesture is like an imaginary sculpture, shaped by the speaker's hands. To give an example of an iconic gesture from this thesis (see page 33): a speaker traced out a large square with her hands in front of her body, to indicate a sign post.

In *pantomimic* gestures, or *enactments*, speakers imitate with their body the acting person the speech refers to. Making a throwing movement when speaking about someone who is throwing is an example of a pan-

tomimic gesture (see page 34 for a more detailed example).

Emblems are "lexicalized" gestures. These gestures have well defined meanings that are shared within a language community. Examples from English and Dutch are the thumb-up "OK" gesture, and the finger-to-the-lip gesture that means "be silent".

Beats are rhythmic up and down motions of the hand that do not seem to represent anything. It is suggested in the literature that beats have rhythmical properties that are related to the phonology of the concurrent speech , but that this relation is more indirect than has previously been suggested (McClave, 1994). Another theory about beats is by McNeill (1992) who claims that beats serve a meta-narrative function, for instance to "underline" words in the speech.

McNeill also defines a class called *metaphoric* gestures. Metaphoric gestures are similar to iconic gestures, but depict abstract objects instead of real ones. The distinction between iconic and metaphoric gestures is not adopted in this thesis. From the viewpoint of language production it makes little difference whether a represented entity is abstract or real, the assumption being that in both cases it will be represented in the speaker's mind, enabling a speaker to gesture about it.

1.2 Current debates

In their attempts to understand the phenomenon of gesture, researchers have focussed on three general issues: *why*, *how*, and *when* do people gesture.

To start with *why*, a heated debate in the literature is going on about what function(s) gesture has. The primary function of deictic gestures and emblems is obviously to communicate something to the listener. For iconic gestures this is less obvious. While Kendon (1994) and also McNeill (1992) argue that iconic gestures are communicative acts, Krauss, Morrel-Samuels, and Colasante (1991), Krauss, Chen, and Chawla (1996), and Rimé and Schiaratura (1991) claim that iconic ges-

tures are not communicative but rather facilitate the process of speaking. In their view speakers gesture for themselves and not for the listener. In chapter 4 this issue will be discussed, and experimental results are presented about the facilitatory effect of gesture on speaking.

One of the arguments used by Krauss et al. (1991) to support their claim that iconic gestures are not communicative is that these gestures are very hard to interpret without access to the accompanying speech. This raises the issue of the "meaning" of iconic gestures. What part of the communicative intention of the speaker (if any) is expressed in gesture, and which part in speech? One way of trying to answer that question is to study what kind of mental representations gestures are generated from. According to Butterworth and Hadar (1989) gestures are generated from lexical entries, i.e. semantic specifications of words. Others (Kendon, 1994; McNeill, 1992) point out that gestures often reveal properties that are not even mentioned in the concurrent speech (see page 77 for a detailed example).

These and other assumptions about the representations involved in gesturing are discussed in chapter 2. In that chapter, the question *how* gestures are produced, in terms of representations and processes operating upon them, is investigated by formulating a hypothetical "blueprint" for a model that incorporates the production of both gesture and speech. Also discussed in that chapter is the *when* issue. Gestures and their accompanying speech are temporally interlocked in apparently systematic, but to a large extent mysterious ways. The temporal dependencies between gesture and speech present an opportunity to test assumptions about the global architecture of the speech/gesture production system. In chapter 3, pointing gestures are shown to be synchronized so tightly with the concurrent speech that one strong assumption made in the model presented in chapter 2 turns out to be wrong.

1.3 Methodologies

To study gesture production, a number of methodologies are available. The most basic methodology is to carefully study video recordings of people who are speaking and gesturing. McNeill (see e.g. McNeill, 1992) often had participants watch a cartoon movie, after which they are required to tell another participant what happened in the movie. Kendon (see e.g. Kendon, 1972) has often studied recordings of conversations "in the wild", which gives a lower degree of control over the content of the speech, but more naturalistic data.

While this phenomenological approach does reveal many ways in which gesture is used in its relation with speech, it has one major disadvantage: for making general claims about the nature of gesture one requires an impractically large amount of data. This is because gesture is a very "noisy" phenomenon. When speakers gesture, and what they gesture about varies considerably, both between and within speakers. A frequently used way to reduce the noise in gesture recordings is to study a number of gestures, classify them into different groups on the basis of some criterion of interest, and then apply inferential statistics to the group counts (see e.g. Kita, 1993).

Once exploratory research leads to well defined hypotheses, it is desirable to run controlled experiments, specifically aimed at testing these hypotheses. However, with iconic gestures and beats it is not possible to use analogues of the standard tasks used in speech production research, such as picture naming or sentence completion. The first problem is that the shape and occurrence of iconic and beat gestures are not governed by shared linguistic rules, as for instance in sign language. This implies that there is a high amount of between-participant and even within-participant variation between gestures that are produced in near-identical contexts. While one can rely on the fact that within a sign language community, signers would all produce the same sign for "chair" when presented with picture of a chair, one cannot rely on speakers of English to produce identical gestures, even when they are describing the same object. A second problem is that beats and iconic gestures seem to be produced only during free and spontaneous speech,

largely without conscious awareness on the part of the speaker. It is theoretically possible to ask participants to make a certain iconic gesture at a certain time, or to produce a sentence accompanied by beat gestures, but the participants would then not display the spontaneous behavior the gesture researcher is interested in. The problem is that experimental control of spontaneous behavior is a contradiction in terms.

For deictic gestures and emblems[2] these experimental limitations do not exist. Pointing and emblems do constitute semiotic systems that are shared within a language community, so there will be little variation between and within participants. Also, participants can point at objects or produce emblems "on command", which makes it possible to ask them to point in controlled experiments. In the experiments reported in chapter 3, this convenient property of deictic gestures is employed to study the temporal synchronization of gesture and speech.

Instead of trying to systematically vary the gestures made by speakers, it is also possible to vary the *conditions* under which they are producing speech and gesture, and then study the properties of the speech and gesture produced under these different conditions. This methodology has mainly been applied in order to answer the question why people gesture (see chapter 4).

1.4 Overview

The remainder of this thesis is organized as follows. In chapter 2 a general architecture for the speech/gesture production system is presented. In this model, called the "Sketch Model", many established findings about gesture are incorporated. It is subsequently compared with two other proposals: one by McNeill (1997), and one by Krauss et al. (1996). It is argued that although McNeill's approach (the theory of growth points) seems to capture the semantic and temporal synchrony of gesture and speech well, it is not specified in terms of rep-

[2] *Mimetics* are a special class of gestures that is also governed by linguistic rules (Kita, 1997).

resentations and processes operating upon them. This makes it almost impossible to derive testable predictions from it. The model by Krauss et al. differs from the Sketch Model in that it adopts the assumption that (iconic) gestures are only performed to facilitate speech, whereas the core assumption of the Sketch Model is that gestures are produced for communicative reasons. These different assumptions about the function of gesture lead to very different processing architectures.

In chapter 3, the temporal synchronization of pointing gestures with the concurrent speech is investigated, using an experimental paradigm in which participants have to name pictures, and point at them as well. The main issue under investigation is how the timing of speech and gesture adapt to each other, both in normal situations and in the case of speech errors.

Chapter 4 is about possible facilitatory effects of gesturing on speech production. Two hypotheses are investigated. The first is that gesturing facilitates speech production by activating the imagistic representations the speech is about, and the second is that gesturing facilitates the process of speech production directly.

The final chapter summarizes and discusses the obtained results. It also suggests possible further research with respect to the issues raised in this thesis.

THE PRODUCTION OF GESTURE AND SPEECH

CHAPTER 2

2.1 Introduction

Research topics in the field of speech-related gesture that have received considerable attention are the function of gesture, its synchronization with speech, and its semiotic properties. While the findings of these studies often have interesting implications for theories about the processing of gesture in the human brain, few studies have addressed this issue within the framework of information processing.

In this chapter, I will present a general processing architecture for gesture production. It can be used as a starting point for investigating the processes and representations involved in gesture and speech. For convenience, I will use the term 'model' when referring to 'processing architecture' throughout this chapter.

Since the use of information processing models is not believed by every gesture researcher to be an appropriate way of investigating gesture (see e.g. McNeill, 1992), I will first argue that information processing models are essential theoretical tools for understanding the processing involved in gesture and speech. I will then proceed to formulate a new model for the production of gesture and speech, called the 'Sketch Mo-

del'. It is an extension of Levelt's (1989) model for speech production. The modifications and additions to Levelt's model are discussed in detail. At the end of the section, the working of the Sketch Model is demonstrated using a number of illustrative gesture/speech fragments as examples.

Subsequently, I will compare the Sketch Model with both McNeill's (1992) growth point theory and with the information processing model by Krauss et al. (1996). While the Sketch Model and the model by Krauss et al. are formulated within the same framework, they are based on fundamentally different assumptions. A comparison between the Sketch Model and growth point theory is hard to make, since growth point theory is not an information processing theory. Nevertheless, the Sketch Model and growth point theory share a number of fundamental assumptions.

An important conclusion is that information processing theories are essential theoretical tools for exploring the processing involved in gesture and speech. The presented Sketch Model is an example of such a tool. It accommodates a broad range of gesture phenomena, and can be used to explain or predict these phenomena within a consistent formal framework.

2.1.1 Terminology

In this chapter, the word 'gesture' is used in the restricted sense of spontaneous body movements that occur during speech and that often appear to represent aspects of the topic of the accompanying speech. Although most gestures are hand gestures, other body parts, such as the head, are also often used for gesture. I will follow McNeill's (1992) typology for distinguishing different kinds of hand gesture. A short overview of this typology (adapted from McNeill (1992)) is presented below:

Iconic gestures: Depicting aspects of the accompanying speech topic.

Metaphoric gestures: Same as iconic, but representing abstract entities.

Deictic gestures: Pointing gestures.

Beat gestures: Biphasic movements of the hands or fingers that do not represent anything.

Emblems: Gestures whose form-meaning relation is lexicalized.

2.2 Information Processing

A common way to formulate theories in cognitive psychology is to use the *information processing approach*. In this approach, theories are often specified by using highly suggestive box and arrow drawings. These 'boxologies' may be helpful visual tools, but they do not reveal the underlying assumptions of the information processing approach. I will therefore clarify and justify these assumptions before presenting a model for gesture and speech.

The term information processing itself is precise: the core assumption of the approach is that the brain does its job by processing information. This is a weak, but highly plausible assumption, supported by a wealth of neuro-biological data. In the formal definition of information by Shannon and Weaver (1949), information is defined to be anything that reduces the uncertainty about a number of possible alternatives. The general nature of this definition is the main reason that the Information Processing assumption is relatively weak. Neural networks, Artificial Intelligence models, and spreading activation models, to name just a few, are all information processing models, even though they differ wildly in the way they represent and process information.

In a stronger version of this approach the word 'information' is taken to mean 'representation'. A representation is some information that is stored as a retrievable entity. The fact that the sun shines is information in Shannon & Weaver's sense of the word, but if I note down in my diary "sun shines today" the entry in my diary will be a representation. The difference between the fact that the sun is shining and my note of that fact is the possibility of retrieving my note at a later time, even when at

that time weather conditions have changed. I will use the abbreviation RP (Representations and Processes) for this approach.

The extra assumption implicit in the RP approach is that we can view representations and the processes operating on those representations as *functionally* distinct entities. From this assumption it does not follow necessarily that processes and representations have different spatial locations in the brain, or that the processes involved operate in a certain order. All it states is that once we have a description of the processes and the representations involved in a certain cognitive activity, we know what computations are performed in order to perform this cognitive activity. Often, the term 'symbolic' or 'classic' is also used to refer to this approach. However, 'classical' theorists usually make stronger assumptions about representations and processes than those of the RP approach (see Fodor & Pylyshyn, 1988).

Even when an RP theory is correct, we have no complete knowledge about the cognitive domain of interest. For example, we do not know how the processes are carried out by our neural hardware, or how the representations are stored in the brain. From the RP perspective, to answer those questions it is necessary to know *what* the brain does before trying to figure out *how* it does it.

Needless to say, it is possible to make mistakes in developing an RP theory. If such a mistake is made, researchers investigating lower levels of processing (e.g. neuro-scientists) can be wrong-footed in their research, because they have an incorrect view of the computation involved. In that case we have no choice but to hypothesize *another* RP theory, and try again. It is impossible to find out how the brain works only by looking 'under the hood'. If there is no understanding of the computations the brain has to perform, even detailed knowledge about the anatomical structure of the brain will hardly be interpretable.

However, it is possible (and desirable) that neuro-scientific knowledge guides and constrains the development of a functional description. For instance, knowledge of the human retina influences our ideas about the kind of representations involved in vision, and neuro-psychological knowledge about human motor control could constrain and inspire

theories about gesture. As Churchland (1986) has argued persuasively, cognitive scientists and neuro-scientists can and should cooperate in order to arrive at a full understanding of the workings of the brain.

Assuming that cognitive faculties can be described by specifying representations and processes at some level of abstraction has many advantages. First, the formal properties of processes operating on representations have been studied extensively by mathematicians and information scientists (e.g., in formal languages and automata). It is possible to use this knowledge in proving certain properties of an information processing model. Second, as with all information processing models, it is often possible to use computer simulations to explore an RP model or parts of it. Simulations are an effective way of checking the coherence of an RP theory. Something important could be missing from an RP theory, or there could be inconsistencies in it. Simulation will reveal such faults, simply because the simulation will either not run or produce the wrong results. This forces the researcher to track down and address the problem. Another advantage of using computer simulations is that the processing assumptions are fully specified in the computer program. Verbal theories or processing accounts often have multiple interpretations, which tends to make them immune to potential falsification.

Many RP theorists make the additional assumption that one or more subprocesses of their model are 'informationally encapsulated' (Fodor, 1983). Models of this kind are often called *modular*. This means, roughly, that computations performed by the subprocess are not affected by computations that take place elsewhere in the system. It should be emphasized that the assumption of informational encapsulation, although it is adopted in the model presented below, does not follow automatically from the assumptions of the RP approach.

Modular models are highly vulnerable to falsification, because they prohibit certain interactions between subprocesses. Any data showing an interaction between two encapsulated processes will be sufficient to falsify the model, or parts of it. Without any modularity, every computation can potentially influence every other computation. This makes both the formulation and experimental falsification of predictions con-

siderably more prone to multiple interpretations. Thus, for any specific case, it makes sense to carry on with a modular RP model until it has been proven beyond reasonable doubt that the modularity assumption is false.

Some researchers believe that making the assumption of modularity is dangerous, for if this assumption is wrong, the knowledge accumulated by means of experimentation can be misleading. For instance, in Levelt's (1989) model of speech production, the phonological representations used by the process of word-form encoding are stored in a lexicon. If lexical retrieval were not a relatively independent process, the knowledge obtained from picture naming experiments could not be generalized to the process of spontaneous speech (S. Duncan, personal communication). However, there is empirical evidence that the results obtained using experimental tasks are comparable to results found under more naturalistic conditions. For instance, the well known effects of semantic priming in reaction time research (see Neely, 1991), have been replicated using the Event Related Potential methodology (Hagoort, Brown, & Swaab, in press), even though participants in E.R.P. experiments typically perform no explicit task – they just passively listen to or read language fragments. Another source of support for modularity is the amazing speed and fluency of speech which makes it likely that there are specialized subprocesses that operate in a highly automated, reflex-like fashion, enabling important subprocesses to operate in parallel (Levelt, 1989, p.2).

McNeill (1987) argues that information processing has certain built-in limitations that make it impossible to apply it to language behavior:

> The most basic [limitation] is that information-processing operations are carried out on signifiers alone, on *contentless* symbols (Fodor, 1980a). Given this limitation the only way to take account of 'meaning' and 'context' is to treat them as inputs that are needed as triggers to get the machinery moving, but that are not modeled by the information processor itself. (McNeill, 1987, p. 133, emphasis in original).

However, the fact that the elements of a computation (symbols) do not have inherent content is not a limitation of information processing theories. As Pylyshyn puts it:

> [Turing's notion of *computation*] provided a reference point for the scientific ideal of a mechanistic process which could be understood without raising the spectre of vital forces or elusive homunculi, but which at the same time was sufficiently rich to cover every conceivable informal notion of mechanism (Pylyshyn, 1979, p.42).

In other words, it is necessary to define computation as operations on form rather than on meaning (or content), for if symbols have inherent meaning, there also needs to be an entity to whom those symbols mean something. This would introduce a homunculus in the theory.

Context information should obviously be incorporated in theories of language processing. The information processing framework allows for that, provided there is sufficient knowledge about the role context plays in speech production.

The only limitation of information processing in general is that it does not allow 'vital forces' or 'homunculi' to be used as explanatory devices. This limitation is in fact one of the main virtues of the approach.

2.3 A model for gesture and speech

The gestures of interest in this chapter usually occur during speaking, and are meaningfully related to the content of the speech. It is therefore plausible that these gestures are initiated by a process that is in some way linked to the speaking process. Sometimes people do gesture without speaking, for instance when speech is not possible (e.g. in a noisy factory), but for the moment I will ignore this phenomenon, to address it later in this section. The fact that gesturing and speaking are in many ways related to each other led to the choice of extending an

existing model for speaking to incorporate gesture processing. Another reason for doing this is to make use of, and be compatible with, existing knowledge about the speaking process. The model of the speaking process that is extended to incorporate gesture processing is Levelt's (1989) model of speech production. I will shortly describe this model here.

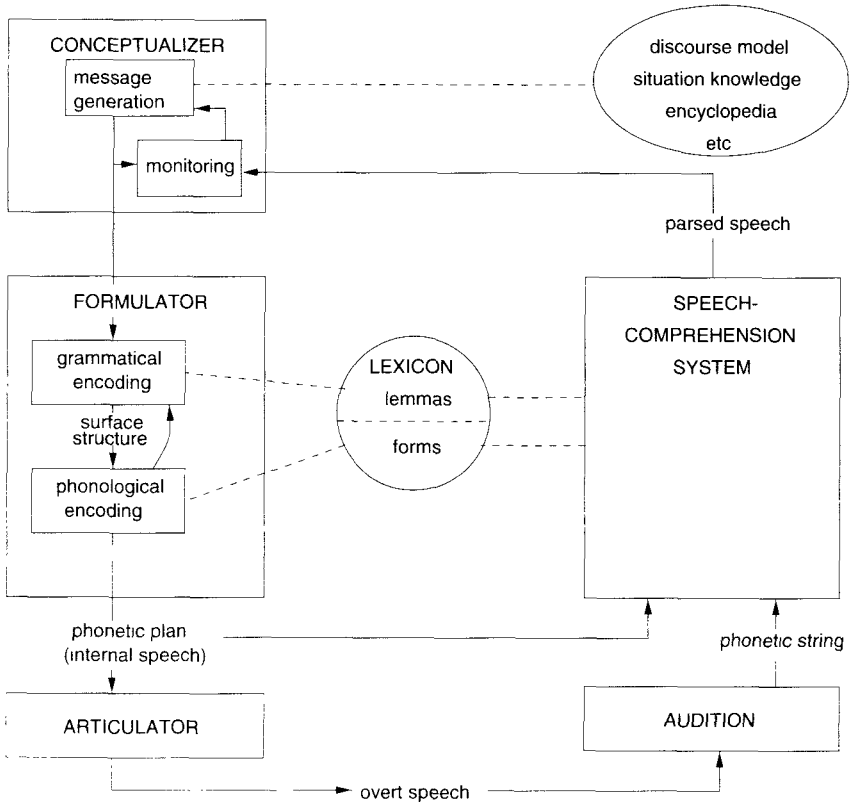


Figure 2.1: Levelt's (1989) architecture for speech production

Figure 2.1 shows the processes (boxes) and knowledge stores[3] (ellipses) defined in the model. In short, given a communicative intention, the

[3]A knowledge store is a collection of retrievable representations.

conceptualizer collects and orders the information needed to realize this intention. It retrieves this information from the general knowledge base represented at the top right of Figure 2.1. The output of the conceptualizer is a representation called the *preverbal message* (or 'message', for short), which contains a propositional representation of the content of the speech. The message is the input for the *formulator*. The formulator will produce an articulatory plan. In order to do that, the first subprocess of the formulator, *grammatical encoding*, will build a (syntactic) *surface structure* that corresponds to the message. It will access a *lexicon* in which the semantic and syntactic properties of lexical items are stored. The second subprocess of the formulator is *phonological encoding*. During phonological encoding, the surface structure built by the grammatical encoder will be transformed into an *articulatory plan* by accessing phonological and morphological representations in the lexicon. The resulting articulatory plan will be sent to the *articulator* which is responsible for the generation of overt speech. Overt speech is available to the comprehension system because the speaker can hear it. Internal speech is also fed back to the speech comprehension system allowing the conceptualizer to monitor internal speech as well[4], and possibly correct it before the overt speech has been fully realized.

In order to extend Levelt's model to incorporate gesture, it is important to make an assumption about the *function* of gesture. Some authors (most notably Kendon, 1994) argue that gesture is a communicative device, whereas others (Krauss et al., 1991; Rimé & Schiaratura, 1991) believe that it is not. There are several arguments for either view. The fact that people gesture when there is no visual contact between speaker and listener (e.g. on the telephone), while this does not present any problems for the listener, is often used as an argument for the non-communicative view. Furthermore, Krauss et al. (1991) argue that iconic gestures (lexical gestures, in their terminology) are hardly interpretable without the accompanying speech. As I have argued in De Ruiter (1995a), there is no real conflict between both views. Ges-

[4]Wheeldon and Levelt (1995) found evidence suggesting that internal speech consists of a syllabified phonological representation

tures may well be intended by the speaker to communicate, and fail to do so in some or even most cases. The fact that people gesture on the telephone is also not necessarily in conflict with the view that gestures are generally intended to be communicative. It is conceivable that people gesture on the telephone because they always gesture while they speak spontaneously - they simply cannot suppress it. Speakers could adapt to the lack of visual contact by producing more explicit spatial information in the verbal channel, but they need not suppress their gesturing. Finally, there is evidence for the view that gesturing facilitates the speaking process (Morrel-Samuels and Krauss (1992), Rimé and Schiaratura (1991), De Ruiter (1995b)), implying that communication could indirectly benefit from the gesturing of the speaker. This could be the reason that speakers do not suppress their gesturing in situations without visual contact.

To conclude, I will assume that gesture is a communicative device from the speaker's point of view. The *effectiveness* of gestural communication is another issue that will not be addressed here.

To extend Levelt's model for speaking to incorporate gesture, the first question to be answered is where gestures originate from. In information processing terminology: What process is responsible for the initiation of a gesture? People do not gesture all the time, nor do they gesture about everything that is being said, so some process must 'decide' whether or not to gesture, and what to gesture about.

Considering the formulation of Levelt's model for speaking, the main candidates within that model for the initiation of gesture are the conceptualizer and the grammatical encoder (the first subprocess of the formulator). Butterworth and Hadar (1989) have suggested that iconic gestures are generated from lexical items. If they are correct, the process of lemma retrieval (a sub-process of grammatical encoding) would be responsible for the initiation of gesture. However, there is ample evidence that most gestures cannot be associated with single lexical items. In a corpus of gestures by native speakers of Dutch (De Ruiter, in preparation), many participants drew a horizontal ellipse in the air with their index finger, while saying "liggend ei", [ENG: *lying egg*]. The gesture did not represent the concept of "lying", nor did it express the concept

"egg". Rather, it represented, simultaneously, the concepts of "lying" and "egg" together. Similarly, in data by the McNeill Gesture Laboratory in Chicago, a participant speaks about a small bird in a cartoon, throwing a bowling ball down into a drain pipe. During the utterance, the participant performs a 'pantomimic' gesture of this action. The gesture reveals aspects of the bowling ball itself, of holding it, of throwing it, and of throwing it in a downwards direction. If gestures were associated with lexical items, one would expect that a gesture only reveals information that is semantically equivalent to the meaning of the 'lexical affiliate'.

Given the fact that many iconic gestures reveal properties that can, at best, only be represented by phrases, the conclusion is that gestures do not have 'lexical affiliates' but rather 'conceptual affiliates'. While it is possible to argue that gestures do have a lexical affiliate, even when there is more information in the gesture than in the affiliate, such a proposal would neither be parsimonious nor empirically supported. There is much evidence, most notably from McNeill (1992) that gestures are synchronized with and meaningfully related to higher level discourse information. The notion of a conceptual affiliate can also explain the occurrence of the occasional gesture that seems to be related to a single word (such as pointing upwards while saying 'up'). All content words have an underlying conceptual representation, but not all conceptual representations have a corresponding content word.

Another reason why the grammatical encoder is an unlikely candidate for initiating gestures is that the formulator's input is a *preverbal message* which is a propositional representation. In other words, it does not have access to imagistic information in working memory. Gestures often represent spatial information that cannot be part of the preverbal message. While it is possible to grant the formulator access to non-propositional information, that would imply a radical change in Levelt's speaking model. My goal is to leave the core assumptions of the speaking model unchanged as much as possible, for a number of reasons. First, there is an extensive body of literature devoted to testing and investigating Levelt's model and its underlying assumptions. This literature will, for a large part, still be relevant to the speech/gesture model

if the speech/gesture model is compatible with it. Second, developing a new speaking model is not only beyond the scope of the present paper but also unnecessary, as I hope to demonstrate below.

Aside from the above considerations, the conceptualizer is the most natural candidate for the initiation of gesture. Many of the problems the conceptualizer has to solve for speech (e.g. perspective taking) also apply to the gesture modality. Furthermore, selecting which information should be expressed in which modality, is a task that is very similar to Levelt's notion of *Macroplanning*, which "... consist in the elaborating of the communicative intention as a sequence of subgoals and the selection of information to be expressed (asserted, questioned, etc.) in order to realize these communicative goals." (Levelt, 1989, p.107).

When people are in a situation that speech is impossible, for instance when they are in a noisy environment, the conceptualizer can choose not to generate speech, but to use gesture to realize the communicative intention. This illustrates once more why the conceptualizer is likely to be responsible for the initiation of gesture. Communicative intentions can be realized in more than one way, and in some cases, gesture can serve to support or replace speech.

Because the conceptualizer has access to *working memory*, it can access both propositional knowledge for the generation of preverbal messages and imagistic (or spatio-temporal) information for the generation of gestures. The conceptualizer will be extended to send a representation called a *sketch* to subsequent processing modules. Because of the central role the sketch plays in the model, I will call it the 'Sketch Model'.

2.3.1 The Conceptualizer

When to gesture

People do not gesture all the time, nor do they gesture about everything they speak about. McNeill (1997) has proposed that gestures occur when new elements are introduced in the discourse. While I have no reason to disagree with this analysis, there are other factors involved

as well. In some cases, it might be necessary to express certain information in gesture, as is most notably the case in pointing gestures that accompany deictic expressions. If someone says "John is over there" without pointing to a location, or when someone says "It is shaped like this", without in some way providing shape information by gesture, the utterance as a whole is semantically incomplete.

Even when gesturing is not obligatory, the conceptualizer can generate a gesture, which would serve the function of enhancing the quality of the communication. The communicative intention is split in two parts: a propositional part that is transformed into a preverbal message, and an imagistic part that is transformed into a sketch.

People are often found to gesture when their speech is failing in some way. Krauss et al. (1991) and Butterworth and Hadar (1989) claim that gesturing in case of a speech failure helps the speech system resolve the problem, for instance by providing the lexical search process with a cross modal accessing cue. Another interpretation, which I prefer, is that the temporary speech failure is recognized in the conceptualizer (e.g. by means of internal or external feedback as in Levelt's model). This recognized speech failure could then be compensated for by the transmission of a larger part of the communicative intention to the gesture modality. Similarly, when circumstances are such that it is difficult to express a communicative intention in speech (e.g. in a noisy environment, or when one does not speak the language) gestures can be generated to compensate for the lack of communicative efficiency in the verbal modality.

The assumption that information that is hard to encode as a preverbal message is encoded in a sketch would also explain why narratives involving salient imagery such as motion events will usually evoke many iconic gestures: The conceptualizer applies the principle that a gesture can be worth a thousand words. As mentioned above, the question whether transmitting information by gesture is always *effective*, i.e. whether the listener will detect and process the information represented in gesture, is another issue.

What to gesture

If imagistic information from working memory has to be expressed in an **iconic** gesture, the shape of the gesture will be largely determined by the content of the imagery. It is the conceptualizer's task to extract the relevant information from a spatio-temporal representation and create a representation that can be transformed into a motor program. The result of this extraction process will be one or more spatio-temporal representations that will be stored in the sketch. Because these representations can involve spatial elements combined with motion, I will call these 'trajectories', for lack of a better term.

Emblems have a lexicalized, hence conventional shape, so they cannot be generated from imagery. I will assume that the conceptualizer has access to a knowledge store that I will call the *gestuary*. In the gestuary, a number of emblematic gesture-shapes are stored, indexed by the concept they represent. If a certain propositional concept is to be expressed, or a certain rhetorical effect intended, the conceptualizer can access the gestuary to see if there is an emblematic gesture available that will do the job. If there is, a reference (e.g. a 'pointer') to this emblematic gesture will be put into the sketch.

A third possibility is that the conceptualizer generates a **pantomimic** gesture. A pantomimic gesture is an *enactment* of a certain movement performed by a person or animate object in the imagistic representation. A good example from the McNeill Gesture Laboratory in Chicago is a tweety-bird narration in which a participant says "and Tweety drops the bowling ball into the drain pipe", while moving both hands as if the participant herself is throwing a bowling ball down. This kind of gesture cannot be generated from imagery alone, but has to be be generated from (procedural) motoric knowledge. The example makes clear why this is the case; Tweety is about half a bowling ball high, and therefore the movement that Tweety makes when throwing the bowling ball is quite different from the movement the (much larger) participant makes in enacting the throwing. For encoding pantomimic gestures, a reference to a motor program (e.g. an action schema (Schmidt, 1975)) will be encoded in the sketch.

Finally, it is also possible to encode a **pointing** gesture into the sketch. This is done by encoding a vector in the direction of the location of the referent. Since the handshape of the pointing gesture is conventionalized (Wilkins, 1995) and for some languages, different types of pointing handshapes indicate the level of proximity of the referent (Wilkins, 1997), the conceptualizer will have to encode a reference to the appropriate pointing template in the gestuary.

As with the production of speech, another important issue that has to be resolved by the conceptualizer is that of *perspective*. If spatio-temporal information is stored in four dimensions (three for space, and one for time), different perspectives can be used to represent the information using gesture. For instance, if a gesture is accompanying a route description (assuming that speaker and listener are facing each other), the gesture that might accompany the speech "take a right turn" might be made to the speakers right (speaker centered perspective) or to the speakers left (listener centered perspective).

In some cultures iconic gestures preserve *absolute* orientation (Haviland, 1993; Levinson, 1996) so in that case the conceptualizer has to specify the orientation of the gesture within an absolute coordinate system. For details about gestural perspectives, see McNeill (1992). A convenient way of encoding perspective in the sketch is by specifying in the sketch the position of the speaker's body relative to the encoded trajectories.

To summarize, the final output of the conceptualizer, called a *sketch*, contains the following information:

Gesture Type	*Sketch Content*
Iconic	One or more spatio-temporal trajectories Location of speaker relative to trajectory
Deictic	Vector Reference to gestuary
Emblem	Reference to gestuary
Pantomime	Reference to motor action schema

The sketch, which will be sent to the *gesture planner*, might contain more than one of the above representations. How multiple sketch entries are processed will be described below.

2.3.2 The Gesture Planner

The gesture planner's task is to build a motor program out of the received sketch. The gesture planner has access to the gestuary, motor procedures (schemata), and information about the environment. One of the reasons a separate module is specified for gesture planning is that the constraints that have to be satisfied in gesturing are radically different from those of manipulating the environment in 'standard' motor behavior.

A problem that the gesture planner has to solve for all gestures, is that of *body part allocation*. One hand may be occupied, in which case either the other hand must be used for the gesture, or the occupied hand must be made available first. If both hands are unavailable, a head gesture can sometimes be generated. For example Levelt, Richardson, and La Heij (1985) had participants point to one of a few lights while saying "this light" or "that light". In one of their conditions participants were not allowed to point. The authors note that

> When no hand gesture is made, the speaker will still direct his gaze or head toward the target LED; there will always be some form of pointing. (Levelt et al., 1985, p. 143)

This illustrates one of the problems the gesture planner has to solve. The sketch specifies a location to be expressed in a deictic gesture, but the hands are not allowed to move. Therefore, the Gesture Planner selects another 'pointing device' to perform the (in this case obligatory) gesture. The fact that the same 'logical' gesture can often be realized overtly by different physical gestures provides support for the assumption that gesture sketch generation and the generation of a motor program are separate processes.

Another task of the gesture planner is to take into account restrictions that objects in the environment impose upon body movements. At a crowded party, too large a gesture could result in hitting another person. Although it probably happens, it seems reasonable to assume that people normally do not hit other people or objects during gesturing. If the environment imposes certain restrictions, the gesture will either be canceled or adapted to fit the circumstances.

For the generation of emblems and pointing gestures, the *gestuary* plays an important role in the creation of a motor program from the sketch. In the gestuary, information is stored about gestural conventions. For instance, while it is possible to point to a location using the elbow or the little finger, in most Western European cultures people point with their index finger. There is a 'soft rule' that specifies that the preferred way to point is to use the index finger of the hand. However, when pointing to a location behind the back, Europeans will usually use their thumb (Calbris, 1990). Speakers of Arrernte (a Central Australian language) will use their index finger for such backward pointings (Wilkins, 1995).

It is not possible to have complete motor programs stored in the gestuary for any of these gesture types. Both pointing gestures and emblems have a number of degrees of freedom. In performing the emblematic "OK" gesture, there is freedom in the location of the hand and the duration of the gesture. The only aspect of this gesture that is fixed is the shape and orientation of the hand. The same holds for pointing: while the shape of the hand is subject to certain conventions, the location that the hand points to is dependent upon where the object of interest happens to be. It is therefore necessary to store gestures in the gestuary in the form of *templates*. A template is an abstract motor program that is specified only insofar it needs to be. Taking the emblematic "OK" gesture as an example, the hand shape is fully specified in the template, while duration, hand (left or right) and location (where the hand is held) are free parameters. The gesture planner will bind these free parameters to the desired values in order to obtain a fully specified motor program.

If the sketch contains one or more trajectories, the gesture planner has to convert them to motor programs that represent the information in these trajectories. In the simple case of one trajectory, the body part (usually

the hand) can often be used to 'trace out' the trajectory, but things can be far more complicated. A more complex problem the gesture planner will have to solve is the generation of so-called *fusions* of different gestures. To take an example from data by Kita (in preparation), a participant is enacting a throwing movement, but adds a directional vector on top of the throwing enactment by 'throwing' the ball sideways, which was not the way the person in the stimulus film threw the ball. However, the movement sideways indicated that (from the speaker's viewpoint) the ball flew off in that particular direction. We must therefore conclude that there has been a fusion of a deictic component (indicating the direction of the ball) and an enactment. This type of fusion of different information in one gesture occurs frequently. If the fusion gesture involves a gestuary entry or an action schema, it is hypothesized that the unbound parameters of the template or action schema (i.e. its degrees of freedom) will be employed, if possible, to encode additional sketch elements. Example C below serves to illustrate this mechanism.

Further research is necessary to find out which types of gesture can be fused together, and under which circumstances fusion of information is likely to occur. Of specific interest is also the question in which order information in the sketch will be encoded in case of a fusion gesture. If, for instance, the sketch contains both a pointing vector and an iconic trajectory, the degrees of freedom left over by the pointing template could be used to represent the iconic component, but the reverse is also possible: the degrees of freedom left over by the iconic gesture might be used to encode a deictic component into the gesture.

Once the gesture planner has finished building a motor program, this program can be sent to lower level motor control units, resulting in overt movement.

To summarize the definition of the gesture planner, upon receiving a sketch, it will:

- generate a gesture from the sketch by retrieving a template from the gestuary (for pointing gestures and emblems), by retrieving an action schema from motoric memory (for pantomimes) or by generating a gesture from the trajectories (for iconic gestures) in

the sketch.

- allocate one or more body parts for execution of the gesture
- try to encode other sketch entries into the gesture as well
- Assess potential physical constraints posed by objects in the environment
- send the motor program to the lower level motor control module(s).

The complete Sketch Model is graphically represented in Figure 2.2. As in Figure 2.1, boxes represent processes, arrows represent representations that are sent from one process to another, ellipses represent knowledge stores and dotted lines represent access to a particular knowledge store.

2.3.3 Synchronization

So far, nothing has been said about the temporal synchronization of gesture and speech. The Sketch Model provides the opportunity to hypothesize in detail how synchronization is achieved.

It should be pointed out that the issue of temporal synchronization is a nebulous one. It is problematic to even define synchronization. The conceptual representation (the 'state of affairs' (Levelt, 1989), or the 'Idea Unit' (McNeill, 1992)) from which the gesture is derived might be overtly realized in speech as a (possibly complex) phrase, and is not necessarily realized overtly as a single word. Therefore, it is by no means straightforward to unambiguously identify the affiliate of a given gesture. Even if a speech fragment has been identified as being the affiliate, it has a certain duration, and so does the gesture. The synchrony between gesture and speech is the synchrony between two time intervals that are often hard to define. However, there is evidence that the *onset* of gesture usually precedes the *onset* of the accompanying speech by a duration of less than a second (e.g. Morrel-Samuels and Krauss (1992),

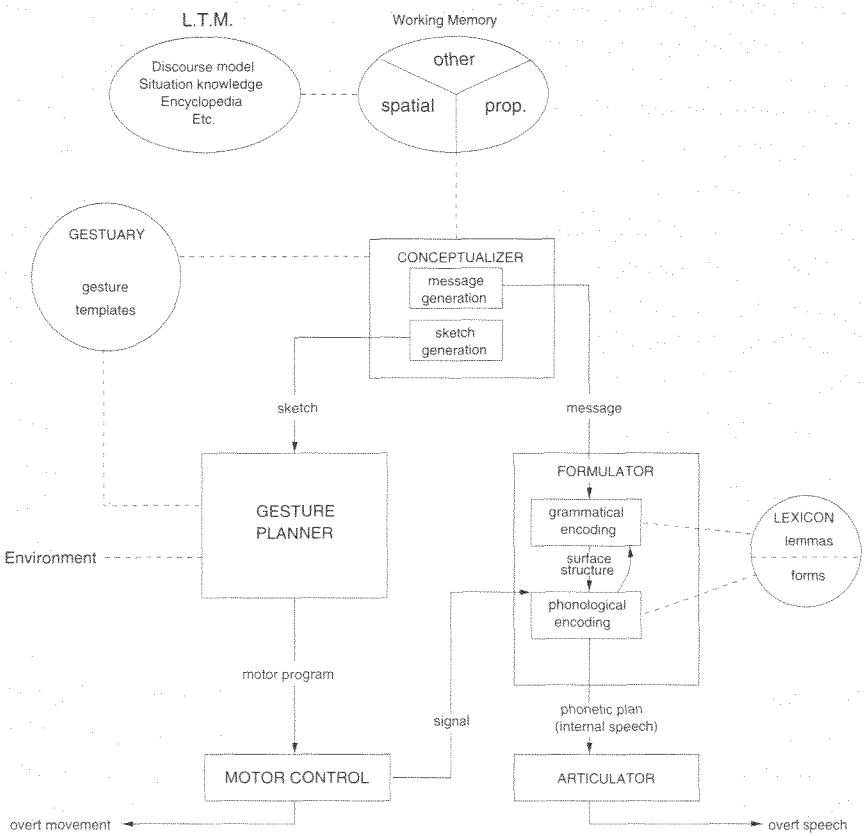


Figure 2.2: The Sketch Model

Butterworth and Beattie (1978), Nobe (1996)). Butterworth and Hadar (1989) conclude that

> ... despite the fact that gestures may continue until after speech onset, the beginnings of movements must be considered as events potentially separable from speech onsets, and any processing model must take this fact into account (Butterworth & Hadar, 1989, p.170).

The Sketch Model can account for this finding by assuming that the generation of gesture takes less time than the generation of speech. Although this assumption by itself is not yet supported by any empirical data, it is a plausible one. Gestures do not have the complicated syntactic properties that spoken language has. Also, in utterances where an iconic gesture is made, the communicative intention often involves imagery. For speech, this imagery has to be translated into propositional format which might require extra processing time.

However, there are other synchronization phenomena that need to be explained. The first to be addressed is the *gestural hold* which comes in two varieties. In the *pre-stroke* hold, the gesturing hand moves towards its initial position, and then waits for the accompanying speech to be produced before performing the stroke (meaningful part) of the gesture. In the *post-stroke* hold, the hand remains motionless after the stroke has been completed, until the related speech has been fully produced. This phenomenon led Kita (1990) to claim that gesture is waiting for the accompanying speech to catch up, in order to achieve semantic synchrony between the gesture and the speech. The pre-stroke hold can be accounted for by assuming that the sketch can be sent to the gesture planner before the construction of the preverbal message has been finished. This allows the gesture planner to prepare the motor program, and send only the part of the program needed to put the hand(s) in initial position to the motor units. [5] When the preverbal message is finally sent to the formulator, the conceptualizer will send a 'resume' signal to

[5]The gesture planner knows which part of the motor program is the meaningful part (stroke) because only the stroke is constructed from the information in the sketch.

the gesture planner, which will then send the rest of the motor program to the motor units. The post-stroke hold can be explained by assuming that the conceptualizer sends a stop-signal to the gesture planner once the production of a preverbal message has been completed (the conceptualizer can detect this because of the feedback loop in the speech production model). This mechanism does not only offer an account for the occurrence of the post-stroke hold, but also for another interesting finding by Kita (1990). He found that repetitive gestures (e.g. a pantomimic gesture for sawing or hammering) do not have post-stroke holds. Instead of the hand stopping at the final position, the repetitive movement is repeated until the related speech fragment has been completed. This difference between repetitive and non-repetitive gestures could be a consequence of the motor programs that are constructed by the gesture planner. For a repetitive gesture, the motor program is specified as a 'loop' – when the stroke has been completed, it starts all over again. Therefore, it will continue until it receives a stop signal. If non-repetitive gestures are completed before the stop signal, there are no motor instructions left, so the hand stays in final position.

So far, the discussed synchronization phenomena did not require the model to specify communication links between processing units 'below' the conceptualizer. However, there are synchronization phenomena that require such a communication.

For deictic expressions accompanied by a pointing gesture, it is possible to unambiguously identify the meaning of the gesture and its relation to the speech. Therefore, synchronizing the pointing movement with the deictic term serves the communicative purpose of establishing a relation between the referent of the pointing and the deictic expression. In the pointing situation, the deictic term (e.g. "this") needs to be synchronized with the pointing gesture, in order to ensure that the listener will be able to connect the gesture information with the speech information correctly. Indeed, there is evidence that deictic expressions accompanied by pointing gestures are tightly synchronized.

Levelt et al. (1985) had participants point at one of a number of lights, while they said "dit lampje" or "dat lampje" (ENG: "this light" or "that light"). They found that the apex (endpoint) of the pointing movement

was synchronized with the onset of the deictic term in speech ("dit" or "dat"), and that this synchronization was achieved by the timing of speech adapting to the timing of the gesture (apex).

It is conceivable that in cases where the interpretation of a certain speech/gesture fragment crucially depends on both the speech and the gesture, as is the case in deictic expressions, synchronization is much more strict than in the case of iconic gestures.

To account for the possibility of tight word-level synchronization of gesture and speech, the Sketch Model allows a synchronizing signal to be sent from the motor execution module to the phonological encoder. Note that a signal from the phonological encoder to the motor execution module would fail to accommodate the finding that speech adapts to gesture and not the other way around (Levelt, Richardson & LaHeij, 1985).

For example, the Sketch Model accounts for the synchronization of deictic expressions and pointing gestures in the following way. The conceptualizer is responsible for selecting the appropriate deictic term: e.g. in English, "this" for proximal, and "that" for distal. If the proximal/distal information is sufficient to single out the referent, no gesture needs to be generated. If it is not, the conceptualizer will send a sketch to the gesture planner, containing the pointing vector. As soon as the formulator reaches the stage of encoding the deictic term phonologically, it stops and waits for a go signal from the motor control unit. In order for the phonological encoder to know when to wait for a go-signal, the formulator should tag the representation of the deictic term in the surface structure. As soon as the motor control unit can approximately predict when the pointing gesture is going to reach the apex, it sends a go-signal to the phonological encoder, which will then encode the deictic term and send it to the articulator.

Levelt et al. also found that if the gesture is physically delayed later than 300 ms. before the apex would normally have occurred, speech cannot adapt anymore. This finding follows naturally from the synchronization mechanism in the Sketch Model. Phonological encoding has an estimated duration of around 300 ms. (Levelt, Schriefers, Vorberg, Meyer,

Pechmann, & Havinga, 1991), so any interruption in the gesture occurring when the phonological encoder is already active cannot change the timing of the speech anymore.

Another phenomenon that could be seen as a kind of synchronization is the finding by Kita (1993) that if speakers interrupt their own speech because they detect an error in it, the gesture is often interrupted simultaneously with the speech. In the sketch model, the interruption of both the speech and the gesture is initiated by the conceptualizer. Upon detection of an error in the speech, the conceptualizer sends a stop signal to the formulator and to the gesture planner. These modules pass on this stop signal to lower processing modules.

If future studies reveal a tighter synchronization between iconic gestures and speech than the model accounts for at the moment, the Sketch Model will have to be adapted to incorporate these findings (and will be, in its present formulation, falsified). However, such studies will have to address a number of problems. First of all, synchronization should be defined in such a way that it is possible to locate the affiliate of any iconic gesture unambiguously. Second, synchronization should be defined carefully. Butterworth and Hadar (1989) pointed out that there are 13 types of temporal relations between two time intervals. Since we can ignore the possibility of two points in time being *exactly* synchronous, we are still left with 6 types of temporal relations from which to choose in defining the meaning of 'synchrony'. Finally, there is a measurement problem. Once the affiliate has been defined, the relevant speech interval can be measured within some degree of accuracy, but for gestures this is much harder. Locating the beginning and end of gestures (even if restricted to the stroke) is often problematic, especially when gestures follow each other rapidly.

It is tempting to interpret synchronization phenomena as evidence for interactive theories of speech/gesture production, in which lower level speech and gesture processing continuously exchange information. To paraphrase Kita (1993), because interactive theories assume that speech processes and gesture processes always have access to each other's internal states, gesture can stop when speech stops. Plausible as this may seem, the explanation is incomplete as long as it does not specify what

information about what processes is shared when in the course of processing. There are multitudes of ways in which speech and gesture processes could share information about their respective internal state. As I hope to have demonstrated in the formulation of the Sketch Model, the computational problems to be solved in the generation of gesture on the one hand, and speech on the other are of an entirely different nature. Therefore, sharing all internal state information all the time is not necessary, and probably amounts to interactive 'overkill'. For the same reason, the assumption in growth point theory (McNeill, 1992) that the generation of gesture and the generation of speech are the same process is suspect. It might be the same *general* process, but then this process must perform two rather different computations: one for gesture and one for speech. That again raises the question what information is shared between these two computations.

2.3.4 Some examples

A few examples will be helpful in illustrating the mechanics of the proposed model. I will illustrate the Sketch Model by explaining what happens in case of (a) an iconic gesture, (b) a pantomimic gesture and (c) an Arrernte pointing gesture[6].

Example A. An iconic gesture. A Dutch participant talks about a Sylvester and Tweety-bird cartoon she just saw. The fragment she is going to describe starts with a big sign saying "bird watchers society".

She says:

"Op den duur zie je zo'n eh, eh, vogelkijkersvereniging ofzo ..."

(After a while one sees a eh eh bird watchers' society or something)

[6]Examples A and B are taken from videotaped narrations collected at the MPI by Sotaro Kita. David Wilkins kindly provided a transcript from his Arrernte field data for example C. Any errors in representing or interpreting these examples are the responsibility of the author.

Roughly during the production of the compound "vogelkijkersvereniging" (ENG: bird watchers' society) the participant uses both index fingers to draw a large rectangle in front of her body.

Computations in the conceptualizer result in the encoding of the introduction of the bird watchers' society in the speech channel. However, the fact that the bird watchers' society was introduced in the cartoon by showing a sign post was encoded in the gesture channel. Therefore, a preverbal message corresponding to "vogelkijkersvereniging ofzo" (ENG: bird watchers' society or something) is sent to the formulator, while a sketch containing the large rectangle is sent to the gesture planner. The hesitation before and after the described gesture/speech fragment was produced suggests that this part of the communicative intention was a separate fragment (or "chunk" as Levelt (1989) calls it). Because the sketch and the preverbal message are sent at the same time, the gesture and the speech are synchronized roughly. Interestingly, the speech does contain enough information to understand the cartoon fragment, but only by observing the gesture the listener can tell that there was a sign involved in the cartoon.

This example also illustrates how the conceptualizer can distribute information over different output modalities.

Example B. A pantomimic gesture. Another participant describing the Sylvester and Tweety-bird cartoon says:

"... enne, da's dus Sylvester die zit met een verrekijker naar de overkant te kijken"

(and eh so that is Sylvester who is watching the other side with binoculars)

During the production of "met een verrekijker naar" (ENG: with binoculars at) the participant raises his hands in front of his eyes as if lifting binoculars to his eyes.

This example illustrates how difficult it can be to establish what the speech affiliate of the gesture is. In this fragment, it is hard to de-

cide whether the gesture corresponds to "met een verrekijker" (ENG: with binoculars), or "zit met een verrekijker naar de overkant te kijken" (ENG: who is watching the other side with binoculars). In the latter case, the synchronization of gesture and speech violates the tendency formulated by Butterworth and Hadar (1989) that gesture onset precedes speech onset. We could therefore assume that the affiliate is the stretch of speech that is synchronized with the gesture. However, this leads inevitably to circular reasoning. Either we infer the affiliate from synchronization, or we infer synchronization from the affiliate. It can't be both at the same time, unless the meaning-relation of a gesture and the accompanying speech can be established unambiguously, as is for instance the case with most deictic gestures. As example A illustrates, gesture and speech can communicate different aspects of the communicative intention, so the affiliate is not necessarily semantically related to the gesture. In case the meaning relation between gesture and speech is not clear-cut, the logic of the Sketch model requires one to infer the affiliate from the synchronized speech, because in the Sketch Model the assumption is made that the sketch and the preverbal message are sent at the same time. In this example the conceptualizer, according to the Sketch Model, encodes a preverbal message corresponding to "met een verrekijker", or possibly "met een verrekijker naar de overkant". At the same time, a sketch is prepared, with a link to the motor schema for holding binoculars. The formulator then processes the preverbal message, and the gesture planner constructs a motor program from the specified motor schema.

Example C. An Arrernte pointing gesture. A (right handed) speaker of Arrernte and Anmatyerre (Central Australia) is holding a baby in her right arm. (Prior to holding the baby she had made a pointing gesture with her right hand.) Now she says:

Ilewerre,	yanhe-thayte,	Anmatyerre
[place-name]	[that/there(mid)-SIDE]	[language/group name]

"(the place called) Ilewerre, on the mid-distant side there,
is Anmatyerre (people's country)."

Roughly during the utterance of "yanhe-thayte" she points to the south-east with her left arm. The hand is spread, and the arm is at an angle of approximately 90 degrees from the body.

Wilkins (personal communication) has observed that the orientation of the palm of the Arrernte spread hand matches the orientation of the surface that is being referred to. To paraphrase one of Wilkins' field notes, if the palm of the hand is facing out and vertical, it could indicate (for instance) paintings spread out over a cliff face.

Such a gesture can be interpreted as a fusion of an iconic gesture (representing the surface orientation) and a conventionalized deictic gesture. This provides support for the hypothesis that the way the gesture planner realizes fusions is to utilize degrees of freedom in gesture templates to represent additional iconic elements represented in the sketch.

Using the Sketch Model, this fragment can be described in the following way. On the basis of geographical knowledge, the conceptualizer chooses the mid-distant proximity for the deictic reference, and encodes it in the preverbal message. The indication of proximity is not sufficient to realize the communicative intention, so the conceptualizer will generate a sketch containing a vector in the direction of the location of the indicated place. The conceptualizer accesses the gestuary to find the appropriate pointing gesture. In Arrernte, both the shape of the hand and the angle of the arm in pointing are meaningful and conventionalized. The pointing gesture that is selected from the gestuary in this example is one with an arm angle of 90 degrees. Given that the indicated place is not visually available from that vantage point, and is at a significant distance away, the 90 degrees angle corresponds with the mid-distant proximity. The spread hand is used to identify a region, something spread out over an area. The entry in the gestuary for this type of pointing gesture specifies hand shape and arm angle, but leaves the parameters specifying the planar angle of the gesture, the 'handedness' of the gesture, and the orientation of the palm of the hand undefined. Finally, the sketch will contain a trajectory that represents a horizontal plane, indicating that the region is spread out over the ground.

The sketch containing the vector, the entry into the gestuary, and the horizontal plane trajectory is sent to the gesture planner. The gesture planner will retrieve the template for the pointing gesture. It will also notice that the right hand is occupied holding a baby, so it will bind the 'handedness' parameter of the gesture template to 'left'. It will bind the parameter specifying the planar angle of the gesture to the angle of the vector that is specified in the sketch, in this case, the south-east. Finally, since the indicated region is a (flat) area, the orientation of the hand will be specified as horizontal. Now all free parameters of the template are specified, yielding a complete motor program to be executed.

2.4 Discussion

To summarize, the most important assumptions of the Sketch Model are:

- The conceptualizer is responsible for the initiation of gesture.
- Iconic gestures are generated from imagistic representations in working memory.
- Gestures are produced in three stages.
    1. The selection of the information that has to be expressed in gesture (the sketch).
    2. The generation of a motor program for an overt gesture.
    3. The execution of the motor program.
- Different gestures can be 'fused' by an incremental utilization of degrees of freedom in the motor program.
- Apart from the conceptualizer, gesture and speech are processed independently and in parallel.

The model covers a large number of gesture types: iconic gestures, metaphoric gestures, pantomimes, emblems, and pointing gestures. The Sketch Model also accounts for a number of important empirical findings about gesture. The semantic synchrony of gesture and speech follows from the fact that both gesture and speech ultimately derive from the same communicative intention. Iconic gestures are derived from imagistic representations, while speech output is generated from propositional representations. The global temporal synchrony of iconic gestures and speech is a consequence of preverbal message (speech) and sketch (gesture) being created at approximately the same moment, while the tight temporal synchrony of (obligatory) deictic gestures is accomplished by a signal from the motor unit to the phonological encoder.

It is interesting to note that in this model, iconic and metaphoric gestures as defined by McNeill (1992) are indistinguishable. Both types of gestures are generated from spatio-temporal representations in working memory. The fact that in the case of a metaphoric gesture the spatio-temporal representation is about abstract entities has no consequence for the transformation of this representation into an overt gesture. On the other hand, pantomimic gestures, while being a subclass of iconic gestures in McNeill's taxonomy, have to be treated in a different way than other iconic gestures, due to the fact that pantomimic gestures can't be generated from an imagistic representation alone, as explained in section 2.3.1.

There is one category of gestures that is not incorporated by the Sketch Model, namely beats. The reason for this omission is that there is, at present, insufficient knowledge available about beats. McNeill (1992) has proposed that beat gestures serve metanarrative functions. Lacking detailed processing accounts for the metanarrative level of speech production, incorporating McNeill's proposal in the model is, at present, not possible. While it is sometimes possible to hypothesize about the role of beats in a given speech/gesture fragment, it is still impossible to *predict* their occurrence using convenient landmarks: "Beats [...] cannot be predicted on the basis of stress, word class, or even vocalization itself." (McClave, 1994, p.65)

Since a major advantage of an information processing model is its vulnerability to potential falsification, it is important to point out a number of predictions that can be derived from the model.

First, the model predicts that people sometimes do *not* gesture. When people read written material aloud[7], or when they are quoting (repeating) someone, they are predicted not to generate spontaneous gestures[8]. In quoting or reading, the conceptualizer is not constructing a new preverbal message and/or sketch. As the Sketch model assumes that the generation of gestures is tightly coupled with the generation of preverbal messages, there will be no spontaneous gestures either.

With respect to the synchronization of iconic gestures with the related speech, the model makes the prediction that the onset of the iconic gesture is (roughly) synchronized with the onset of the overt realization of the conceptual affiliate in speech (usually a noun phrase or a verb phrase), independent of the syntax of the language. For example, if an English speaker says "he runs across the street", the onset of the iconic gesture that represents "running across something" is predicted to be roughly co-occurring with the onset of the first word of the verb phrase, in this case "runs". If a Japanese speaker says "miti o wata te" (street go-across) the onset of the iconic gesture is predicted to be synchronized with the onset of "miti" (street); although it has a different meaning than the first word in the English sentence, it is again the first word of the verb phrase [9]. The reason the model predicts synchronization with the verb *phrase* and not with the verb is that the gesture is assumed to be planned together with its corresponding preverbal message, which is assumed to have NP's and VP's as minimal "chunks".

Another prediction concerning the synchronization of gesture and speech is that the model does not permit that representations active in the formulator will influence the timing of the gesture. Lexical stress, or pitch accent, for instance, are therefore predicted to have no effect on gesture/speech synchronization.

[7] Assuming their hands are free

[8] Of course, it is conceivable that people "quote" the gestures made by the quotee, but these gestures are not spontaneous.

[9] I took these example sentences from McNeill (1992).

Finally, the word level synchronization mechanism that accounts for the synchronization of obligatory pointing gestures is predicted to be operative for any gesture that is semantically obligatory. If someone says: "the fish I caught was this big", making a gesture indicating the size of the fish, the gesture is predicted to be just as tightly synchronized with the word "this" as is the case with obligatory pointing gestures.

2.4.1 A comparison with growth point theory

In comparing the Sketch Model with McNeill's (1992) growth point (GP) theory, it is possible to point out both similarities and differences. The main similarity is that according to both the Sketch Model and GP theory, gesture and speech originate from the same representation. In the Sketch Model this is the communicative intention, while in GP theory it is the growth point. "The growth point is the speaker's minimal idea unit that can develop into a full utterance together with a gesture" (McNeill, 1992).

McNeill (1992) discusses and dismisses theoretical alternatives (pp. 30–35) for GP theory. His analysis applies equally well to the Sketch Model, because of two important assumptions that underly both GP theory and the gesture model: gestures and speech are part of the same communicative intention, and are planned by the same process.

However, GP theory does not give any account of how (in terms of processing) growth points develop into overt gestures and speech. The growth point is an entity whose existence and properties are inferred from careful analyses of gestures, speech fragments, and their relation (McNeill, 1992, p. 220). However, without a theory of how a GP develops into gesture and speech, gesture and speech data can neither support nor contradict GP theory. This also introduces the risk of circularity. If a fragment of speech that is accompanied by a gesture is interpreted as a growth point, and the growth point is also the entity responsible for the observed (semantic and temporal) synchronization, the observation and the explanation are identical. Therefore, GP theory in its present form does not explain how the speech/gesture system actually accomplishes

the observed synchrony.

2.4.2 A comparison with the model of Krauss, Chen & Chawla (1996)

Figure 2.4.2 shows the information processing architecture proposed by Krauss et al. (1996), which is also based upon Levelt's (1989) model[10]. For convenience, I will call their model the KCC model. The most important difference with the Sketch Model is that in the KCC model the conceptualizer from Levelt's model is left unchanged, whereas in the Sketch Model it is modified extensively. In the KCC model, gestures are not generated by the conceptualizer, but by a separate process called the Spatial/Dynamic Feature Selector. These features are transformed into a motor program which helps the grammatical encoder retrieve the correct lemma for the speech. Contrary to the Sketch Model, the model of Krauss et al. does incorporate the generation of beats (motor movements in the Krauss et al. terminology). In their model, prosodic specifications generated by the phonological encoder are translated into beat gestures. Synchronization is accounted for in the KCC model by the assumption that the phonological encoder can terminate gestures by sending a signal to the motor planner unit once the affiliated lexical item has been encoded phonologically.

The main assumption of the KCC model is that iconic gestures (lexical gestures, in their terminology) are not part of the communicative intention, but serve to facilitate lemma selection. However, if the spatio-dynamic features in the generated gesture are to facilitate lemma selection, it is essential that the features that, taken together, single out a particular lemma are identifiably present in the motoric realization of the gesture. If the gesture contains features that are not associated with the lemma that is to be retrieved, these features will very likely *confuse* (hence slow down) lemma selection, because more than one lemma will be activated by the features in the gesture. The example given by

[10] It should be noted that most of the arrows in the model by Krauss et al. have a different meaning that the ones used in both the Sketch Model and Levelt's model.

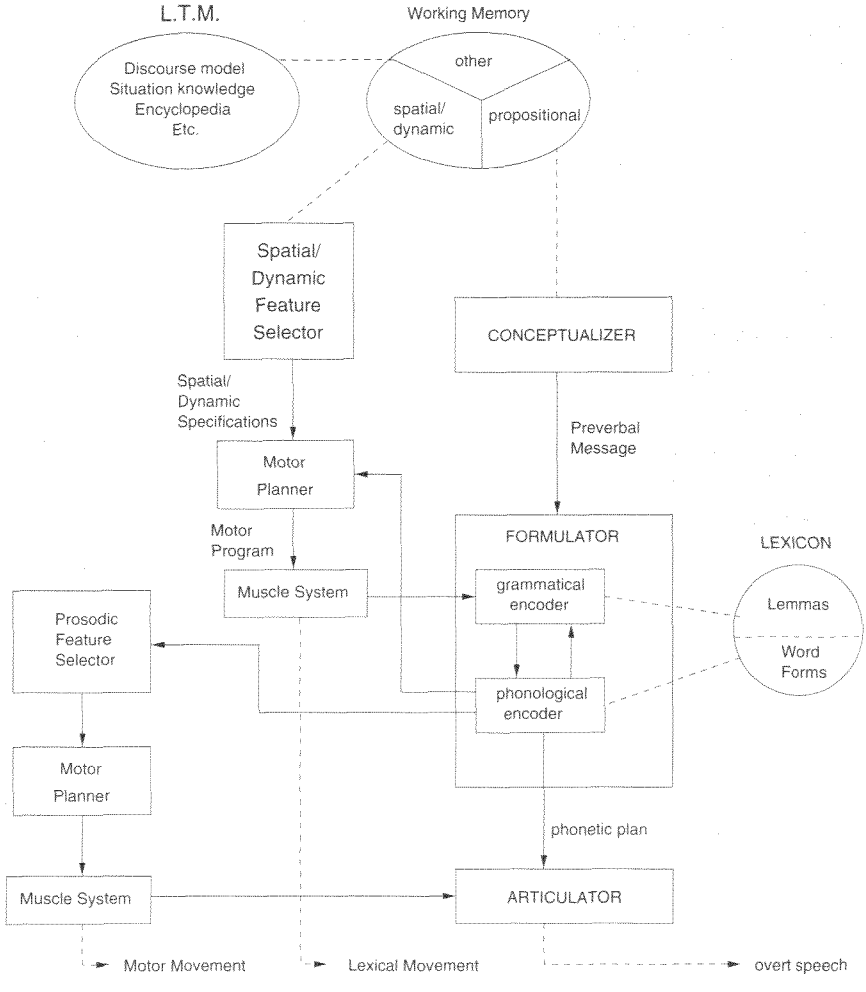


Figure 2.3: The model by Krauss, Chen & Chawla (1996)

Krauss et al. is a gesture about a vortex. A gesture representing a vortex might indeed contain features that, taken together, would help retrieving the lemma for "vortex". Interestingly, a gesture containing these features will be very easy to identify as representing a vortex, even by a listener who does not have access to the accompanying speech. Especially when, as Krauss et al. assume, the features present in the gesture are to be apprehended proprioceptively, these features must be clearly present in the overt realization of the gesture.

This is in contradiction with the experimental findings reported in Krauss et al. that indicate that most iconic or lexical gestures are not easily recognizable at all without access to the accompanying speech. The paradox here is that those gestures that could facilitate lemma selection must be gestures whose meaning is largely unambiguous (such as the gesture for "vortex"). Gestures that are hard to interpret without the accompanying speech will usually not contain enough information to facilitate lexical selection.

Another problem with the KCC model is that if gestures are indeed derived from spatio-dynamic working memory (an assumption their model shares with the Sketch Model), there is no reason to expect that there exists a single lemma that has the same meaning as the gesture. As mentioned above, in many cases *phrases* are needed to describe or approximate the meaning of a gesture. Therefore, if gesturing facilitates the speaking process, this is more likely to involve more complex (higher level) representations than lemma's. For instance, De Ruiter (1995b) found evidence suggesting that performing iconic gestures facilitates the retrieval of imagery from working memory (see also chapter 4).

Synchronization in the KCC model also differs from that of the Sketch model. While the mechanism Krauss et al. propose is more parsimonious than that of the Sketch Model, it is doubtful whether it will explain all synchronization phenomena. Especially problematic for the KCC model is the post- and pre-stroke hold phenomenon. In the KCC model gestures are terminated once the corresponding lexical item has been retrieved, but especially the pre-stroke hold phenomenon indicates that gestures can also be *initiated* in synchronization with speech.

2.5 Conclusions

In investigating the processing involved in gesture and speech production, it is a great advantage to develop theories within the framework of information processing. IP theories are less prone to multiple interpretations, easier to falsify, and they facilitate the generation of new research questions. These advantages become even more salient when the theory is in a stage of development that allows for computer simulations. The IP approach is a theoretically neutral formalism that enhances the resolution of our theories, without restricting their content, apart from the fact that the IP approach does not allow the use of 'homunculi' as explanatory devices.

The Sketch Model is an attempt to incorporate and explain many well established findings in the area of gesture and speech with a largely modular IP model. It incorporates a large amount of knowledge about gesture and speech within a consistent framework. As has been shown, predictions can be generated from the Sketch Model, that can be tested in future experiments. The Sketch Model allows for detailed specification of hypotheses about the synchronization between gesture and speech. It can also accommodate cross-cultural variation in gesture behavior, and proposes a new and detailed account for the fusion of information in gestures. Because the Sketch Model is an extension of Levelt's (1989) model, the model implicitly incorporates a large amount of accumulated knowledge about speech production.

While the formalism used to specify the Sketch Model is similar to the formalism used by Krauss et al. (1996), the underlying assumptions of the Sketch Model are more similar to growth point theory. Most notably, the assumption that gestures and speech are planned together by the same process and ultimately derive from the same representation is made in both the Sketch Model and in growth point theory. However, contrary to the Sketch Model, growth point theory does not specify how a growth point develops into overt speech and gesture.

WHEN'S THE POINT?

THE SYNCHRONIZATION OF POINTING AND SPEAKING IN NON-DEICTIC EXPRESSIONS

CHAPTER 3

3.1 Introduction

It has often been observed that spontaneous hand gestures accompanying speech are synchronized (i.e., are produced at approximately the same moment in time) with the fragment of speech that is meaningfully related to the gesture. Little is known, however, about the mechanisms responsible for this synchronization during the production of gesture and speech.

Kendon (1980) has formulated what McNeill (1992) later called the *phonological synchrony rule*. This rule states that the stroke (meaningful part) of a gesture precedes or ends at, but does not follow the phonological peak syllable of speech. Kendon presented evidence from recorded conversations to substantiate his claim. Nobe (1996) investigated and confirmed the phonological synchrony rule using six videotaped narrations of a cartoon movie. The present study is aimed at investigating speech/gesture synchronization using pointing gestures in an experimental setting.

Pointing gestures have distinct advantages for investigating synchro-

nization. They have a predictable shape (within a certain language community), and the affiliate can usually be identified unambiguously, especially in *deictic* pointing gestures.[11] Levelt et al. (1985) defined a deictic gesture to be a gesture without which a deictic utterance is incomplete. If one says "look over there" without some form of pointing gesture, head nod, or other kind of paralinguistic indication of where "there" is, the utterance is essentially incomplete. A deictic gesture is therefore *obligatory*. Deictic pointing gestures also have a well defined semantics: they can be replaced by speech that has the same function. One could replace "look over there", accompanied by a pointing gesture to the north east by the sentence "look in the north-east direction" and the semantics will be, roughly speaking, the same.

Levelt et al. investigated the synchronization of speech and deictic pointing gestures by having their participants look at a horizontal array of LEDs. When one of the LEDs flashed, their participants had to indicate which LED flashed by saying "this light" or "that light" and pointing at it. By measuring both speech onset and the motion trajectory of the pointing hand, Levelt et al. could show that (a) the apex[12] of the pointing was synchronized with the onset of the deictic word ("this" or "that"), (b) this synchronization was accomplished by speech adapting to the timing of the gesture, and (c) gesture execution was ballistic - once the gesture has started to execute, speech processes have almost no influence on the execution of the gesture anymore.

In deictic expressions involving obligatory gestures this high degree of synchronization is functional; it is essential that a listener establishes a connection between the deictic element in the speech and the information provided by gesture. Iconic gestures, on the other hand, are generally not essential for understanding speech (Krauss et al., 1991), which can be illustrated by the fact that people often produce iconic gestures during telephone conversations.

The present study investigates the synchronization of *non-obligatory*

[11] I will use the word *deictic* to refer to the function of the gesture, and the word *pointing gesture* to refer to the shape of the gesture.

[12] Levelt et al. defined the "apex" of the gesture to be the point in time when the arm is maximally extended, and the hand comes to a near stop for a short period of time.

pointing gestures. A non-obligatory pointing gesture is a pointing gesture that is not essential for understanding the speech, but does provide extra information. If someone says "Could you give me the pencil?", it is possible that "the pencil" provides the listener with sufficient information to single out the intended object. However, if the speaker points to the pencil as well, the listener is informed about the location of the intended object, which helps the listener to locate it. Although non-deictic pointing gestures are generally not identical to iconic gestures, they share their non-obligatory nature, which could lead to another type of synchronization than Levelt et. al observed with deictic (and therefore linguistically obligatory) pointing gestures.

Kendon (1980) has claimed that gestures are synchronized with the peak syllable of the phonological phrase. Assuming that by this peak syllable is meant the (last) pitch accented syllable in an intonational phrase, and given that the pitch accent will go to the primary stressed syllable within the word (henceforth "lexically stressed syllable"), the first hypothesis is that the lexically stressed syllable will provide the synchronization point of the gesture in one-word utterances.

3.2 Experiment 1

The objective of this experiment is to investigate whether the apex of a non-obligatory pointing gesture is synchronized with the production of lexically stressed syllables in the accompanying speech. An operationalization of the concept of "stroke" (meaningful part) of a pointing gesture is therefore necessary. As in Levelt et al. (1985), the *apex* of the pointing gesture is assumed to be the most meaningful part of the pointing gesture, for it is at the apex that the index finger is actually aimed at the target.

3.2.1 Method

Participants

A total of twelve native Dutch speakers participated in the experiment. There were three male and nine female participants. All participants were right handed. All participants reported to have normal vision, and were paid for their services.

Procedure

The participants were seated at a table. At the edge of the table opposite the participant, a plate of transparent matted plexiglass was mounted vertically. Between the edge of the table and the plexiglass there were four red LEDs attached, with 47 cm between the LEDs (see Figure 3.1). Between the two middle LEDs, an additional LED (henceforth called the "fixation LED") indicated the fixation point. Four different line drawings were projected on the plexiglass from behind the table using two slide projectors. Each of the four pictures was projected above one of the LEDs (excluding the fixation LED). The projectors were placed such that the brightness of the four pictures was the same, seen from the participant's position. The room was darkened to enable the participants to see both the projected drawings and the LEDs clearly. The projected pictures were white lines on a black background with a white frame around them of 18 × 18 cm. The participants were seated at the table, and had to put their right hand on a hand-shaped piece of rubber on the table. Under the table, right below the fixation LED, a speaker was placed for the warning tone.

The participants were instructed to look at the fixation LED, and keep their right hand flat on the hand-shaped piece of rubber between trials. A trial started with a 330 Hz warning tone of 500 ms, during which the fixation LED was on. After another 1000 ms, one of the four LEDs below the pictures flashed for a duration of 1000 ms. The participant was instructed to name the picture above the LED that flashed, in the format [definite determiner][noun] (e.g. "de camera", [ENG: *the camera*])

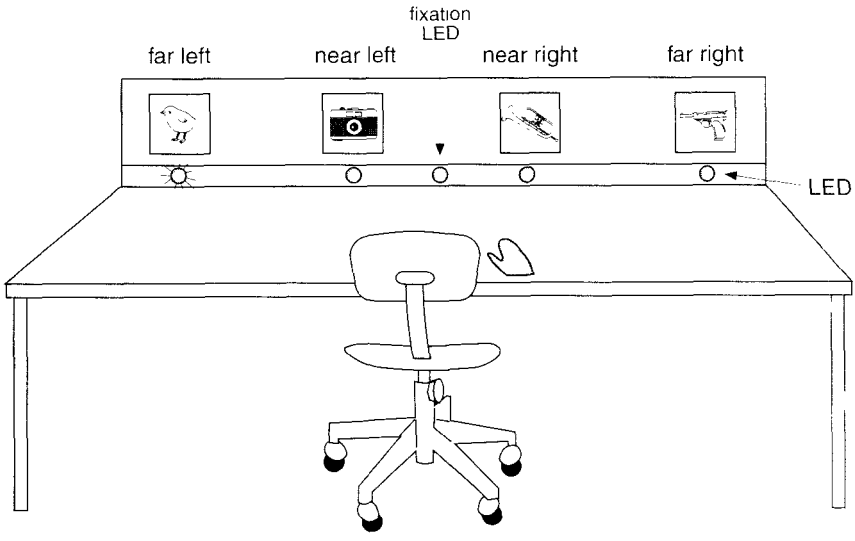


Figure 3.1: Experimental setup

while pointing at that picture. The participants had 2500 ms from the moment the LED flashed to complete a response. After having completed a response, they were requested to put their hand back on the hand-shaped piece of rubber. The instruction did not emphasize speed. Before the actual experiment, participants were familiarized with the pictures used in the experiment, and their preferred names. The experiment started with 22 practice trials (with other pictures than the ones used in the actual experiment) to familiarize the participants with the task.

Note that the instruction used the Dutch equivalent of the phrase "name the picture while pointing at it" as well as "point at the picture while naming it". This does give the participants the instruction to point and speak at roughly the same time, but it does not suggest in any way how speech and gesture are to be synchronized at microtiming level.

Apparatus

Speech was recorded using a microphone attached to a cord around the neck of the participant, which was connected to the left channel of a DAT recorder. On the right channel of the DAT recorder, short high frequency pulses were recorded that enabled time alignment with the motion data. The motion data were collected with a Zebris(TM) CMS 50 motion analyzer. A small ultrasound buzzer was attached to the left side of the index finger knuckle of the right hand. The ultrasound signal this buzzer generates was received by the Zebris(TM) receiver unit consisting of three microphones in a plane, placed vertically at the left side of the table. This setup enabled motion registration with a temporal resolution of 5 ms (200 Hz sampling rate) and a spatial resolution of 1 mm. The motion data were low-pass filtered using "Kernschätz" filtering, which is a filtering method similar to finite impulse response filtering (Marquardt & Mai, 1994). The filtering windows used were 50 ms for the position, 70 ms for the velocity and 90 ms for the acceleration. After data collection, the speech signal and the motion data were synchronized to enable time locked display using a waveform editor.

Materials and design

A total of 16 pictures were used. The pictures were chosen such that the initial phoneme of the stressed syllable of the picture name would be a (possibly voiced) plosive or a sibilant, to make it easier to localize the onset of these phonemes using a waveform editor. Also, the picture names were all monomorphemic. Eight of the pictures had bisyllabic names and another eight had trisyllabic names. These groups each consisted of four pictures of which the name had lexical stress on the initial syllable, and four with lexical stress on the final syllable. Finally, these groups each consisted of two words with neuter grammatical gender (in which case the definite determiner is "het") and two words with common grammatical gender (definite determiner "de"). The difference in gender was added to the design to make the task more natural, because in natural speech the grammatical gender of the noun has to be retrieved

in order to use the correct definite article.

Four pictures were presented simultaneously in the four different positions on the projection screen (see Picture 3.1). From left to right, these positions are called Far Left, Near Left, Near Right, and Far Right. A group of four pictures presented at the same time will be referred to as "picture group". For each picture group, six trials were run, such that two pictures occurred as a target once, and two other pictures twice, resulting in six trials per picture group. Having six trials per picture group was necessary in order to present all four positions at least once, without participants being able to guess which position would come up in the next trial.

All pictures appeared at least once as a target in all four picture positions. There was a total of 16 picture groups (four pictures presented simultaneously) in the experiment, which resulted in a total of 96 (16×6) trials. Every picture group had two bisyllabic and two trisyllabic names, and also two names with neuter and two with common grammatical gender. The picture groups were arranged such that every combination of gender, stress position, number of syllables and pictures occurring twice as a target within a picture group were all occurring equally frequently for each position. The picture groups were subsequently randomized, with the constraint that picture groups that contained the same four pictures were presented consecutively, to avoid the participants having to adjust to a new set of names for every picture group. The resulting trial order was presented in reverse order (with respect to both the presentation of the picture groups and the trial order within a picture group) to half of the participants in order to counteract possible effects of trial order. See appendix A and B for a complete description of the stimuli, and the composition of the picture groups.

Results

Of the twelve participants that participated in the experiment, four were excluded from the analysis. One participant could not complete the experiment due to an equipment failure during the experiment. Two par-

ticipants, in spite of being encouraged to point clearly, lifted only their index finger, which resulted in null measurements because the motion sensor was attached to the hand, and not the finger itself. A final participant was removed because he completed the entire pointing movement (including the returning of the hand to the table) and only then gave a verbal response. This pointing/speech behavior was atypical and also resulted in timeouts, because the speech was produced while the next trial had already begun. 21 trials (2.7%) were marked as an error because the participants either did not respond within 2500 ms after the LED flashed, or because they could not complete the response within the available 2500 ms. These errors probably occurred because the participant did not see the LED flash, or saw it too late.

The dependent variables with regard to the motion of the hand are BP (begin pointing), and AP (apex). The speech related dependent variables are BA (beginning of definite article), EA (end of definite article), BN (beginning of noun), BS (beginning of stressed syllable), ES (end of stressed syllable), and EN (end of noun). BP was calculated by locating the time sample in which the absolute velocity of the hand exceeded 1% of the maximum velocity reached in that particular trial. The apex was defined as the sample where the hand reached its maximal forward extension. The dependent variables related to the speech signal were all localized using a waveform editor, by using both auditory and visual inspection of the signal. An overview of the relative timing data is given in Figure 3.2. Speech timing is represented by bars above the time lines, and gesture timing by the indicators below the time lines.

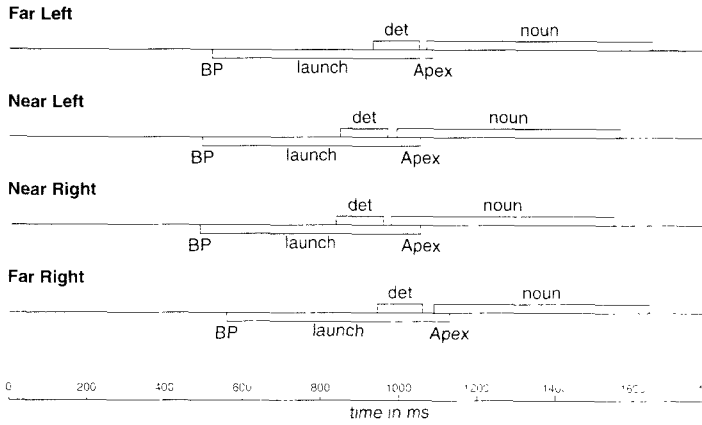


Figure 3.2: Timing overview of Experiment 1 (by picture position)

Due to the large number of dependent variables, in the subsequent analysis only effects that are either directly relevant to the discussion, or are significant at the 5% level in the participant analysis, the item analysis, or both, will be discussed.

Considering first the gesture data, pointing to the two peripheral pictures (henceforth called FAR pictures) was, on average, initiated 50 ms later than pointing to the two central (NEAR) pictures ($F_1(1,7) = 10.86$, $MS_e = 6555$, $p = .013$, $F_2(1,15) = 9.24$, $MS_e = 1891$, $p = .008$). The apex occurred on average 57 ms later when pointing to peripheral targets. The 7 ms difference in total pointing duration (AP − BP, henceforth called LAUNCH) for the FAR and NEAR positions was not significant ($F_1(1,7) = 1.25$, $MS_e = 4580$, $p = .3$, $F_2(1,15) = 2.98$, $MS_e = 393$, $p = .11$).

An odd effect in the data is an interaction between the factor stress and picture position in the analysis of the pointing initiation times (BP). For the Far Left picture position, pointing was initiated 64 ms earlier when the lexical stress was word-final than when it was word-initial. Note that this effect is in a direction opposite to the one predicted by the phonological synchrony rule. This difference between the initial

and final stress items is only significant for the Far Left picture position ($F_1(1,7) = 24.0$, $MS_e = 719$, $p = .002$, $F_2(1,14) = 4.84$, $MS_e = 3370$, $p = .045$), and not for the other positions, and the interaction between the factors picture position and stress is significant ($F_1(3,21) = 5.87$,$MS_e = 1266$, $p = .004$, $F_2(3,24) = 6.63$, $MS_e = 1166$, $p = .002$). Since this effect only occurred for the leftmost picture position, and not for the rightmost position, it cannot be due to the fact that pointing to a FAR picture takes more time than to a NEAR picture. The effect was probably caused by some list order effect that was not counteracted by the reversal of the list for half of the participants.

The finding by Levelt et. al (1985) that speech adapted to gesture was replicated here, as is shown by an analysis of BA (onset of definite article, or voice onset). BA is 95 ms later when FAR pictures are named ($F_1(1,7) = 16.75$, $MS_e = 20224$, $p = .005$, $F_2(1,15) = 42.53$, $MS_e = 1869$, $p = .005$). This could, however, just be a reflection of the pointing *and* the speech being delayed in the FAR trials, for instance because perceiving FAR pictures takes more time than perceiving NEAR ones. However, Levelt et al. (1985) found that naming-only latencies were not affected by the location of the target, a finding that was replicated by Feyereisen (1997), providing strong evidence against the hypothesis that the speech was delayed only because of perceptual effects.

An important question now is: what is it that the speech/gesture system attempts to synchronize with the apex? It is clear from Figure 3.2 that the apex is *on average* temporally very close to noun onset, but that might only be the case for the average value. Therefore, a regression analysis was performed, predicting apex times using a number of "landmarks" in the speech. Table 3.1 below gives the temporal distances between the landmarks and the apex, as well as the β coefficients from the regression analysis. The β coefficient is an indicator of the strength of the relation between the predictor (in this case a speech landmark) and the dependent variable (here the apex). The column "Sign." indicates the statistical significance of the proportion of variance explained by the independent variable.

As is apparent from Table 3.1, the best predictor of the apex times (AP)

Table 3.1: Apex Distances and Regression Weights

Landmark	Apex dist.	β	Sign.
Begin article	-189	.27	< .001
Begin noun	-48	.41	< .001
Begin stressed syllable	70	.03	.45
End stressed syllable	357	.01	.73
End noun	520	.11	.004

is the onset of the noun (BN). The variables BS and ES (onset and offset of the stressed syllable) have no predictive power with respect to the apex timing: they have low and nonsignificant β coefficients in the regression equation.

Another way to assess the effect of stress location on apex times is to look at the apex in both stress conditions. If the location of the stressed syllable has any influence on the time course of the gesture at all, it should be expected that the apex occurs later when the stress is word-final. This is not the case. In the initial stress group, the average apex time is 1088 ms, while in the final stress group it is 1077 ms, which is in the wrong direction, and not significant (F_1 and F_2 both < 1).

There is, however, the oddball effect of stress for the leftmost picture position mentioned above. This effect could potentially suppress an effect of stressed syllable position, because for that position, pointing occurred earlier for the picture names with final lexical stress. Therefore, the analyses were redone, excluding the trials where the picture position was Far Left. The results are presented in Table 3.2 below.

Excluding the Far Left picture position from the analyses shows an even more marked role of BN: *only* the onset of the noun is predictive, and it is more predictive than in the previous analysis. BN by itself explains 65% of the variance in the apex timings. The average apex time in the initial and final stress conditions were 1080 ms and 1084 ms, respectively, which is not significant ($F_1(1,7) < 1$, $F_2(1,14) = 3.27$, $MS_e = 2899$, $p = .092$).

Table 3.2: Apex Distances and Regression Weights without Far Left Position

Landmark	Apex dist.	β	Sign.
Begin article	-202	-.12	.25
Begin noun	-59	.81	< .001
Begin stressed syllable	71	.07	.11
End stressed syllable	360	.04	.25
End noun	506	.06	.11

3.2.2 Discussion

To summarize the findings, Experiment 1 replicates the findings of Levelt et. al. that speech adapts to gesture, which suggests that this synchronization mechanism is not dependent on the obligatory nature of the pointing gestures in their experiment. In addition, the pointing gestures are synchronized with the onset of the noun.

The location of the lexically stressed syllable does not have any influence on the timing of the gesture, but on a strict interpretation, the phonological synchrony is not violated in Experiment 1. The apex does indeed always precede the strong syllable. However, under this strict interpretation, the phonological synchrony rule is more a kind of *constraint* than a synchronization principle. Therefore, in the remainder of this study the phonological synchrony rule is interpreted as also predicting that the timing of the stroke *covaries* with the temporal location of the strong syllable. This version of the phonological synchrony rule will be called the *strict* phonological synchrony rule. If the only strong syllable in a fragment of speech can be considered the "peak syllable of the phonological phrase", the strict phonological synchrony rule is falsified by the present results. However, the noun phrases produced in Experiment 1 do have a relatively sparse intonational structure, so it is not clear whether the stressed syllables in Experiment 1 are really "peak syllables".

Therefore, in the next experiment, *contrastive stress* (or *accent*) is used to further investigate the strict phonological synchrony rule.

3.3 Experiment 2

The most prominent syllables in an utterance are the pitch accented syllables. At least one of these will be there in any utterance, but frequently more than one word in the utterance will be provided with a pitch accent. When a word has what is known as contrastive stress, this means that its pitch accent stands out even more, because the other words in the utterance are de-accented (i.e. are deprived of their pitch accents) while the remaining contrastive pitch accent may be pronounced with an expanded pitch range (Shattuck-Hufnagel & Turk, 1996). For example, in the sentence "No, she's not my wife, she's my SISter", the word "sister", establishes a contrast with "wife", and is therefore emphasized. This contrast need not be with another word in the sentence, but can also be related to the discourse context. If one sees ten butterflies, and only one of them is green, saying "the GREEN butterfly" emphasizes the color name because it is the color that enables the listener to single out the intended one.

In Experiment 2, participants produced sentences such as "the GREEN crocodile" (emphasizing the color name) or "the green CROcodile" (emphasizing the object name). The usage of the [defdet][adj][noun] format, in combination with the contrastive stress on either the adjective or the noun makes it more likely that the strict phonological synchrony rule, if it indeed exists, will have an effect. The new design enhances the phonetic realization of stress, it allows for a wider range of stressed syllable positions, and it introduces a more marked intonational contour in the production of the speech.

3.3.1 Method

Participants

A total of eleven native Dutch speakers participated in the experiment. There were two male and nine female participants. All participants were right handed. All participants reported to have normal vision, and were paid for their services.

Procedure

The experimental setup was the same as in Experiment 1, except for the instruction, the timing, and the pictures used. Participants either saw four different objects in the same color, or the same object in four different colors. The participants were instructed to emphasize that element that was distinctive within the context of the four pictures that were presented at the same time. If they saw four crocodiles in different colors, and the LED would flash under the green crocodile, they were instructed to say "the GREEN crocodile", emphasizing "green". If, on the other hand, they saw four different objects, all in the color green, while the LED flashed under the (green) crocodile, they were instructed to say "the green CROcodile", emphasizing "crocodile". As in Experiment 1, participants were instructed to point to the indicated object while they described it.

Before the experiment, the participants were familiarized with the pictures used in the experiment, and their preferred names. Participants were instructed to look at the fixation LED between trials. A trial started with a 330 Hz warning tone of 500 ms, during which the fixation LED burned. After another 1000 ms, one of the four LEDs flashed for a duration of 2000 ms. From the moment the LED flashed, participants had 4000 ms to complete a response, after which the next trial would start. The experiment started with 24 practice trials (4 picture groups, 6 trials per picture group) to familiarize the participants with the task.

Materials and design

Of the four monomorphemic trisyllabic object names used in the experiment two had word-initial stress ("camera", [ENG: *camera*] and "boterham", [ENG: *slice of bread*]), and two had word-final stress ("krokodil", [ENG: *crocodile*] and "hagedis", [ENG: *lizard*]). These were all names with common grammatical gender, because systematically varying the grammatical gender of the noun in this experiment would have led to an unacceptably long experiment.

Unfortunately, there are no polysyllabic color names in Dutch with word initial stress. The color names used were therefore either monosyllabic ("geel", [ENG: *yellow*] and "groen", [ENG: *green*]) or trisyllabic with word-final stress ("violet", [ENG: *violet*] and "antraciet", [ENG: *a blueish shade of grey*]).

For each of the four colors, four picture groups were made, featuring all four objects names, but in varying positions. Similarly, for each of the four object names, four picture groups were made that had the four different colors at varying positions. Again, as in Experiment 1, one picture group involved six trials, to minimize the predictability of the LED position. The picture groups were arranged such that every combination of contrastiveness (adjective vs. noun), stress position within the contrastive word (initial vs. final), and pictures occurring twice as a target in a picture group occurred equally frequent for every position. The picture groups were subsequently randomized. The resulting trial sequence was presented in reverse order (with respect to both the presentation of the picture groups and the trial order within a picture groups) to half of the participants in order to counteract possible effects of trial order.

Results

Pre-analysis. Of the eleven participants, three had to be excluded from the analysis. One participant made pointing movements by only lifting the finger (which resulted in null measurements due to the lo-

cation of the buzzer on the hand), one participant could not detect the LEDs during fixation on the fixation LED, and one participant did not complete the experiment due to lack of time.

The pre-analysis of the data is similar to that of the previous experiment, with the following differences. First, more speech landmarks had to be located in the speech stream. The variables that have been located for each trial are BU (beginning of utterance), BA (beginning of adjective), EA (end of adjective), BN (beginning of noun), EN (end of noun). In addition, the beginning and end of the stressed syllables of both the adjective (BAS and EAS) and the noun (BNS and ENS) were located.

With respect to the motion data, an intriguing difference with the previous experiment was that participants did not immediately retract their hand after it reached the maximal extension, but held it out for a notable duration.

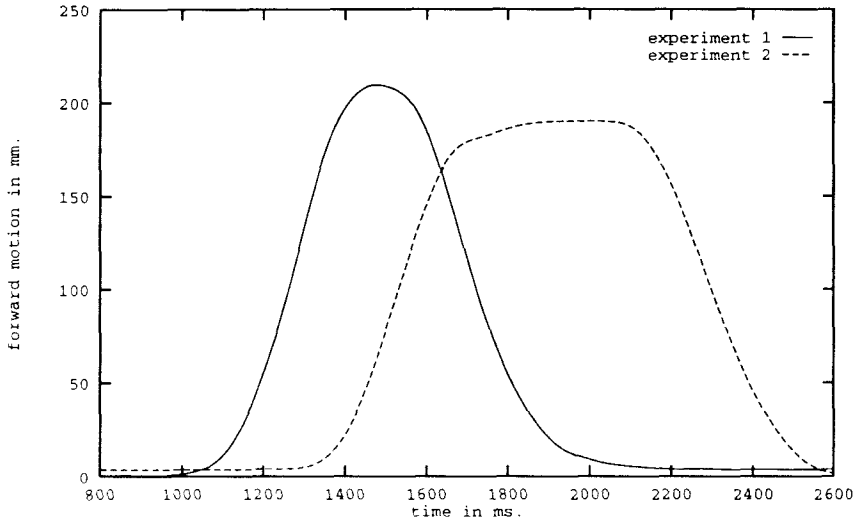


Figure 3.3: Typical motion patterns for exp. 1 and 2

In Figure 3.3, a motion plot is shown of a typical trial from both Experiment 1 and Experiment 2. Time is represented on the horizontal axis, and the extension of the hand (in mm) is represented on the vertical axis.

It is clear that in Experiment 2, the apex is not a very short moment, as in Experiment 1, but has a substantial duration. Therefore, instead of having one dependent variable for the apex, two values were located: BAP (beginning of apex) and EAP (end of apex). BAP is defined as the data sample in which the forward extension of the hand exceeds 95% of the maximum extension reached during the entire trial. EAP is the first sample (past BAP) that the hand is extended less than 95% of the maximum extension.

Different types of error-trials were also excluded from the statistical analysis. The total number of errors was 61 (4%). In 28 trials (1.8%) the participants' speech hesitated or was interrupted and repaired, in 11 trials (0.7%) the contrastive stress was put on the wrong word (e.g. putting contrastive stress on the adjective instead of the noun), in 8 trials (0.5%) the location of the accent could not be determined from the speech signal, in 9 trials (0.6%) participants used the wrong color name (e.g. "paars", [ENG: *purple*]) instead of "violet", [ENG: *violet*]). Finally, in one trial the participant did not complete the response within the available amount of time, and in another trial the participant started a response before the LED flashed.

Speech Adapting to Gesture. As in Experiment 1, speech timing adapts to the timing of the gesture. In Figure 3.5 an overview is presented of the timing data for each picture position. For the FAR picture positions, pointing was initiated 60 ms later than for the NEAR pictures ($F_1(1,7) = 18.27$, $MS_e = .3337$, $p = .004$, $F_2(1,15) = 20.50$, $MS_e = .1245$, $p < .001$).[13] The beginning of the apex (BAP) was 79 ms later for the FAR positions ($F_1(1,7) = 11.43$, $MS_e = .9339$, $p = .012$, $F_2(1,15) = 30.77$, $MS_e = .1537$, $p < .001$), while for the end of the apex (EAP) the difference was 83 ms ($F_1(1,7) = 15.65$, $MS_e = .7618$, $p = .005$, $F_2(1,15) = 26.43$, $MS_e = .2223$, $p < .001$). For the FAR positions speech was initiated 90 ms later than for the NEAR positions ($F_1(1,7) = 31.57$, $MS_e = .4379$, $p < .001$, $F_2(1,15) = 37.78$, $MS_e = .1519$, $p < .001$).

[13] In this experiment, item analyses (F_2s) were performed on the 16 noun phrases, because they corresponded with the 16 different types of pictures that were used.

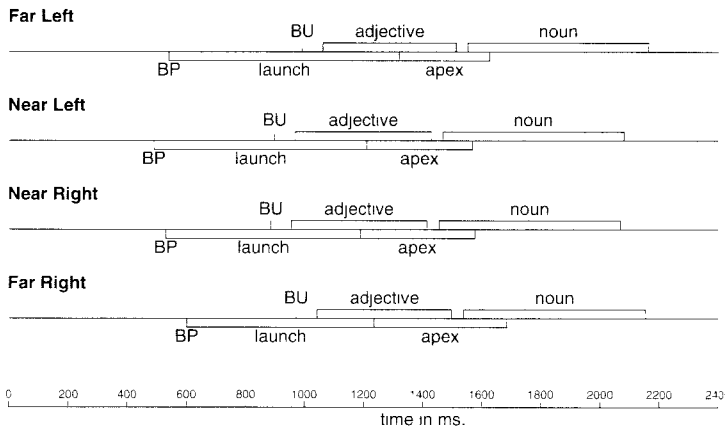


Figure 3.4: Timing overview by picture position

The duration of the total utterance (i.e. the entire NP) hardly varied with the distance of the picture position. The average NP duration was 1182 ms for all picture positions except the far left position, where it was 1172. A post-hoc analysis (Student-Newman-Keuls, $\alpha = .05$) revealed that the difference between the far left position and the other three positions was significant by participant but not by item, suggesting that the 10 ms deviation for the far left position was caused by only a few items.

The speech-to-gesture adaptation could have been the result of a learning process that occurred during the experiment. To test this, the correlation between the gesture-speech asynchrony (GSA = BAP - BU) and the position of the item in the experimental sequence was computed over all trials. This correlation is positive and not significantly different from zero ($r = -.05$, $p = .07$). If the participants would gradually learn during the experiment to adapt their speech timing to the duration of the pointing, GSA would have tended to become smaller with every trial, resulting in a negative correlation. Therefore, it can be concluded that the speech adaptation to the duration of the gesture is not a result of learning during the experiment.

Gesture Adapting to Speech. With respect to the adaptation of the timing of gesture to properties of the associated speech, it should be noted that although contrastive stress (adjective vs. noun, henceforth called CONTRAST) and the lexical stress within the word with contrastive stress (initial vs. final, henceforth called STRESS) are orthogonally varied in the experimental design, they can nevertheless not be analyzed as independent factors. First, the two kinds of adjectives used are monosyllabic color names (the "initial stress" adjectives) and trisyllabic color names with final stress (the "final stress adjectives") on the other. This was unavoidable since Dutch does not have polysyllabic color names with initial stress. The nouns in the materials were all trisyllabic, with either word-initial or word-final lexical stress. In other words, the STRESS distinction is a different one for adjectives than for nouns. Second, CONTRAST and STRESS are not independent, because the actual temporal location of the stressed syllable does not only depend on whether the word in question is initially or finally stressed, but also on whether it is the adjective or the noun that is contrastive.

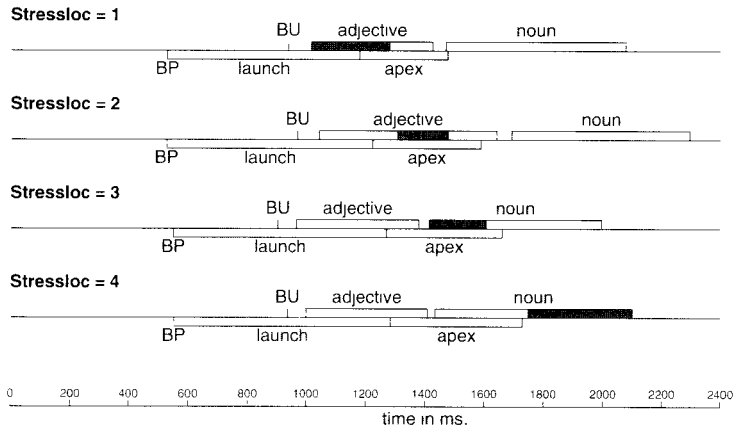
Table 3.3: Stress Location Mapping and Average Values

CONTRAST	*STRESS*	*Loc.*	*Onset*	*Offset*
Adj	Init	1	1015	1280
Adj	(pre)Final	2	1307	1477
Noun	Init	3	1415	1605
Noun	Final	4	1750	2100

For these reasons, the factors STRESS and CONTRAST are collapsed in the analysis to one factor indicating the location of the stressed syllable. In Table 3.3, the mapping of the factors CONTRAST and STRESS to the four different stress locations is defined. Also in this table are the averaged temporal locations of the stressed syllables in the speech data.

An overview of the timing data for every stress location (averaged over all picture positions) can be found in Figure 3.6 below. The dark bars indicate the temporal location of the stressed syllables in the speech. The reason that the stressed syllables in adjectives with final stress are not at the end of the word is that in Dutch, adjectives are inflected when they

are preceded by definite articles. For example, the root form /vi-o-lEt/ then becomes /vi-o-lE-tə/.



Loc.	BP	**Launch**	BAP	**Apex Dur.**	EAP
1	530	**648**	1177	**299**	1476
2	528	**695**	1223	**365**	1588
3	551	**719**	1270	**391**	1661
4	553	**731**	1284	**446**	1730

Figure 3.6: Timing data by stress location (durations in boldface)

The first analysis concerns motion onset (BP). There is no main effect of stress location on BP ($F_1(3,21) = 1.88$, $MS_e = 1803$, $p = .163$)[14]. However, it appears that BP is sensitive to the location of the contrastive element in the planned speech (see Figure 3.6 below). For stress locations 1 and 2, in which the contrastive element is the adjective, BP is 530 ms and 528 ms respectively. For stress locations 3 and 4, where the contrastive element is the noun, BP is 551 ms and 553 ms respectively. Pointing is initiated 23 ms earlier when the contrastive element appears earlier in the produced speech, independent of where the lexical

[14]Because stress location is dependent on the nature of the item *and* on the value of CONTRAST, no by-item analysis could be performed for this factor.

stress of the contrastive element is. This effect is marginally significant in the ANOVA ($F_2(1,7) = 4.59$, $MS_e = 1057$, $p = .069$, $F_1(1,15) = 3.17$, $MS_e = 1949$, $p = .095$), but a one tailed t-test does reach conventional levels of significance ($t_1(7) = -2.05$, $p = .04$, $t_2(15) = -1.86$, $p = .04$, one tailed).

Moving forward in time, the next dependent variable of interest is the duration of the outward motion of the hand, which is computed as BAP - BP (the LAUNCH). It is clear from Figure 3.6 that LAUNCH is longer when the location of the stressed syllable occurs later in the speech. The main effect of stress location is significant ($F_1(3,21) = 9.89$, $MS_e = 2284$, $p < .001$). Post-hoc analyses (Student-Newman-Keuls, $\alpha = .05$) show that only for stress location 1, the mean of LAUNCH is significantly shorter than for stress locations 2, 3 and 4, and that the means for locations 2, 3, and 4 do not significantly differ from each other.

Also, within the group of trials where the contrastive stress was on the adjective, LAUNCH is 48 ms shorter when STRESS is initial, than when STRESS is final ($F_1(1,7) = 95.53$, $MS_e = 96$, $p < .001$). However, the factor STRESS did not have a significant effect on LAUNCH for trials in which CONTRAST = Noun ($F_1(1,7) < 1$).

In other words, only when the contrastive stress is on the adjective does the location of the stressed syllable have an effect on the duration of the outward motion of the pointing hand.

There is, however, an alternative explanation of the effect of the location of the stressed syllable on the pointing times that needs to be investigated. It is conceivable that at a low level of motor processing both pointing and articulating are "locked" to a certain degree, in the sense that during the articulation of a stressed syllable, the pointing motion is sped up. This simple "locking" phenomenon could not only explain the observed (weak) synchronization with the stressed syllable, but it could also explain why the effect of the location of the stressed syllable on the pointing duration is significant only between stress location 1 and stress locations 2, 3 and 4: only for stress location 1 is there a substantial overlap in time between the articulation of the stressed syllable

and the forward motion of the hand (see Figure 3.6). If this motor-locking indeed exists, there should be a negative correlation between the amount of overlap between the forward motion of the hand and the articulation of the stressed syllable (henceforth called OVERLAP) on the one hand, and the duration of the forward motion (LAUNCH) on the other. The longer the OVERLAP, the more the pointing motion should be sped up, thereby shortening LAUNCH. However, the correlation between LAUNCH and OVERLAP is only .009 ($p = .744$), indicating that there probably is no such motor-locking involved. However, this correlation might be low because for stress positions 3 and 4 there is hardly any overlap between LAUNCH and the location of the stressed syllable. Therefore, the same correlation was computed using only trials in which stress position is 1 or 2. This correlation is also low and not significantly different from zero ($r = .03$, $p = .48$).

After the pointing hand has started to move (BP) and completed its outward motion (LAUNCH), it reaches the apex. As explained before, in this experiment the apex has a certain duration, so two measures of apex have been coded: BAP (begin apex) and EAP (end apex). As can be seen in Figure 3.6, due to the combined effects of BP and LAUNCH, BAP increases with stress location. The main effect of stress location on BAP is significant ($F_1(3,21) = 7.32$, $p = .002$). Post-hoc analysis (Student-Newman-Keuls, $\alpha = .05$) reveals that for stress position 1, BAP was only significantly earlier than BAP in stress positions 3 and 4. However, the correlation of BAP with the actual location of stressed syllable onset (computed over all trials) was .61 ($p < .001$). This correlation shows that the phonological synchrony rule is confirmed even in the strict interpretation.

A final variable of interest is the duration of the apex (EAP - BAP), henceforth called APEXDUR. From Figure 3.6 it is clear that APEXDUR also increases with stress location. The main effect of stress location is significant ($F_1(3,21) = 22.64$, $p < .001$). Post-hoc analyses (Student-Newman-Keuls, $\alpha = .05$) reveal that APEXDUR for location 1 is significantly shorter than APEXDUR in locations 2, 3, and 4, and that for both locations 2 and 3 APEXDUR is significantly shorter than for location 4.

Speech Error Analysis. As mentioned before, participants produced a number of erroneous responses. 11 of those errors were interruptions (participants interrupted their own speech and repaired it) and 17 were hesitations (participants hesitated between words). These 28 errors were produced by six different participants. Most errors occurred because participants had trouble with the color names. For instance, they had to say "antraciet", [ENG: *a blueish shade of grey*] instead of the more natural "grijs", [ENG: *grey*], and "violet", [ENG: *violet*] instead of the more natural "paars", [ENG: *purple*]. An error-inducing factor for the nouns was the fact that the pictures of the lizard and that of the crocodile looked somewhat similar. Since on error trials, participants still pointed in an apparently normal way, these error trials were analyzed separately, for they might reveal what happens with synchronization when the speech stream is interrupted. For each of the error-trials, the following values were coded: the start of the speech (BU), the start of the hesitation or interruption, the resumption of the speech (either after an interruption or a hesitation), and the pointing variables (BP, BAP, EAP). Furthermore, average values for identical or near-identical trials *without* an error were extracted from the data from the same participant. For 12 error trials, it was possible to use the data from an identical trial, because those trials occurred twice within a picture group, and the other occurrence was a correct trial. For the remaining 16 error trials, the average was taken of all correct trials with the same picture position, adjective, and the lexical stress position of both the adjective and the noun (so only the noun token was different). This enabled a direct, within-participant and within-trial comparison between error trials and (near-)identical correct trials. These data are summarized in Table 3.4. The variable INTERD in the table represents the duration of the interruption or hesitation. The column "Significance" indicates the (two-tailed) significance level in a paired t-test of the observed mean difference.

Interestingly, in error trials speech started on average 166 ms later than in correct trials, even though the interruption or hesitation itself was only noticeable *after* the articulation of the definite article (see Figure 3.7 for a visual presentation of the error data). If the selection of color names in the early planning of speech is an activation/selection mech-

Table 3.4: Relative Timing of Errors and Correct Trials (averages)

	Correct	Error	Δ	Significance
BU	1021	1187	166	$p < .001$
BP	628	695	67	$p < .05$
LAUNCH	691	808	117	$p < .01$
INTERD	0	211	211	n.a.
APEXDUR	420	624	204	$p < .001$

anism along the lines of Roelofs (1992) this delay can be explained by assuming that when the wrong color name is active (e.g. "purple") the correct one ("violet") will be active too, which results in competition between the two color names, delaying the final selection of the (wrong) color name. More interesting, however, is that the timing of the pointing is adjusted to compensate for the delay in speech onset. The pointing initiation (BP) is delayed by 67 ms, and the duration of the forward motion (LAUNCH) is prolonged by 117 ms, resulting in a total delay of the onset of the apex (BAP) of 184 ms. This almost restores the temporal distance between BAP and BU to the same value as in the correct trials. However, this could be true only for the means, and not for individual trials. Therefore, the correlation was computed between a variable called ADJUST (= Δ BP + Δ LAUNCH) and Δ BU (the delay in speech onset). The correlation between ADJUST and Δ BU was .78 ($p < .001$), which suggests that this temporal adjustment is a systematic phenomenon occurring on the level of individual trials. A similar phenomenon was observed for the apex, which appears to get longer by approximately the amount of the duration of the hesitation or interruption (see Figure 3.7). The correlation between Δ APEXDUR and INTERD was .75 ($p < .001$) for hesitations, and .70 ($p < .05$) for interruptions, suggesting that the apex was lengthened to compensate for the extra speech time taken up by the interruption or hesitation. This account is supported by Kita's (1993) corpus of iconic gesture accompanying speech errors. He found that iconic gestures that did not end before or at the moment speech was repaired either had an extended post-stroke hold or, in the case of repetitive gestures, an extended repe-

tition duration.

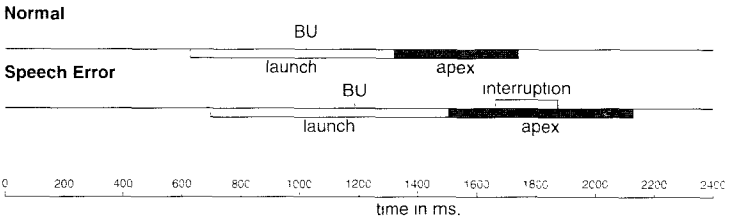


Figure 3.7: Timing of hesitations and interruptions

3.3.2 Discussion

To summarize the findings, as in Experiment 1, the timing of the speech adapts to the timing of the gesture. Furthermore, the duration of the entire produced noun phrase was hardly affected by the timing of the pointing. Motion onset was sensitive to which element of the noun phrase had contrastive stress — if the contrastive stress was on the adjective, pointing was initiated 23 ms earlier than when it was on the noun. The location of the stressed syllable did not have any effect on motion onset. It did, however, have an effect on the pointing duration time. The duration of the forward motion increases with stress location, although the only significant difference was found between stress location 1 on the one hand, and stress locations 2, 3, and 4 on the other. The resulting apex onset times show a similar pattern. The duration of the apex shows a more robust effect of stress position, such that the later the stressed syllable occurs, the longer the apex lasts.

These results support the strict phonological synchrony rule, under the assumption that the beginning of the apex is the most meaningful component of a pointing gesture. Apex onset always precedes the peak syllable, and it is positively correlated with the temporal location of this syllable. However, even though the pointing gestures in the experiment

are sensitive to the location of the peak syllable, there is no *complete* synchronization between the peak syllable and the pointing: the peak syllables are not articulated simultaneously with the apex. Perhaps there is an attempt at synchronizing, but it is not enough to obtain full synchrony. This might be a consequence of the experimental setting. The instruction did not in any way emphasize speed, but the general pace and rhythm of the experiment might have induced a strategy of starting to point as soon as the LED was detected. If this is actually the case, the partial synchronization effects might have been the result of a competition between the strategy of starting to point as quickly as possible and the strict phonological synchrony rule. This issue could only be resolved by having experimental data in which participants have more freedom in their timing, or by analyzing a sufficient amount of naturalistic data.

3.4 General Discussion

The phonological synchrony rule is strongly supported by the results from Experiment 2, even when it is interpreted as predicting that the gesture timing covaries with the timing of the peak syllable. The robust null effects in Experiment 1 with regard to the influence of the stressed syllable suggest that a real intonational (as opposed to only metrical) contour is necessary for the strict phonological synchrony rule to operate as predicted.

The findings from especially Experiment 2 reveal interesting properties of the synchronization of gesture and speech. In order to interpret these findings in terms of information processing, it is necessary to make assumptions about the processing architecture involved in the simultaneous production of gesture and speech. In De Ruiter (to appear) (see chapter 2) such a tentative speech/gesture architecture is outlined. This architecture is based on Levelt's (1989) speaking architecture (see Figure 3.9 for a simplified overview of the model). In this model, when the *conceptualizer* sends a semantic representation of the speech to the *formulator*, it can also send an abstract representation of a gesture to a

gesture planning module. In other words, the conceptualizer initiates the production of a gesture at the same time as it sends the affiliated speech to the formulator.

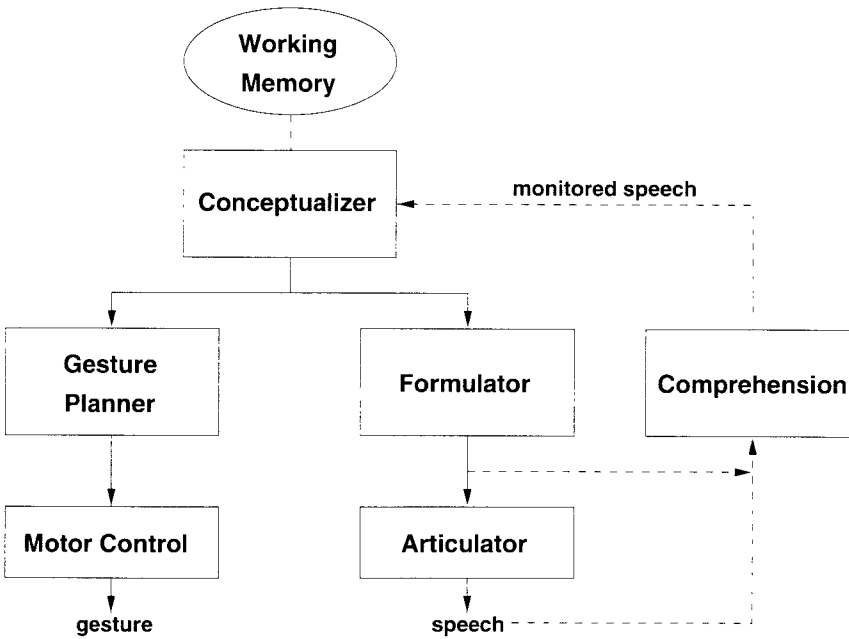


Figure 3.9: Gesture and speech production architecture

A number of assumptions in the model that are relevant here are (1) gestures are initiated by the conceptualizer, but (2) the conceptualizer does not specify the timing of the gesture. (3) There is, at least for non-obligatory gestures, no interprocess communication between modules "below" the conceptualizer. Finally, (4) once the gesture stroke has been completed, the gesturing hand(s) will remain at stroke-final position — the so-called post-stroke hold phenomenon (Kita, 1990) — until the conceptualizer receives the feedback (by means of the speech monitoring loop) that the affiliated fragment of speech has been completed.

A number of assumptions of the proposed architecture are supported by

the data from this study. First, the effect of the accented word on the initiation of pointing in Experiment 2, in combination with the absence of any effect of the location of the lexically stressed syllable on pointing initiation time, suggests that indeed the conceptualizer initiates the gesture; under the assumptions of the architecture (and of Levelt's speaking model) the conceptualizer assigns contrastive stress to the contrastive element in the preverbal message, but it does not have access to phonological information such as lexical stress. Similarly, the fact that in Experiment 2 the duration of the NP was hardly affected by the picture location suggests that the produced NP was delayed (by the conceptualizer), but that the interword timing within the NP is not influenced by the formulator or articulator. This supports the claim by Levelt et al. (1985) that the synchrony between gesture and speech is established during the planning phase, and not during articulation itself.

Finally, the lengthening of the apex, either in case of a speech error or when the location of the peak syllable occurred later in the speech, is compatible with the assumption that post-stroke holds are maintained until the feedback loop provides the information that the affiliate of the gesture has been produced successfully. This would explain why only the duration of the apex is reliably sensitive to the location of the peak syllable, while earlier landmarks (apex onset, or launch) are only marginally affected. If the operation of the strict phonological synchrony rule is by itself not sufficient to obtain full synchronization, the conceptualizer might compensate by lengthening the gesture upon receiving feedback from the comprehension system. The idea that the conceptualizer "holds" the gesture until it receives feedback is also supported by the difference in apex duration between Experiment 1 and Experiment 2 (see Fig. 3.3). In Experiment 1, the duration of the speech was shorter, in which case the conceptualizer could well have received the feedback that the speech was produced *before* the gesture was completed. In that case, the hand can be retracted immediately after reaching the apex, as also happened with the short speech fragments used in the experiments by Levelt et al. (1985). In Experiment 2, it is much more likely that a gesture reached its apex before feedback information arrives at the conceptualizer, leading to a noticeable apex lengthening,

or post-stroke hold[15]

The model is also falsified in two important respects. The fact that speech adapts to gesture, even for these non-obligatory pointing gestures[16], implies that speech onset must have been delayed in pointing to peripheral pictures, relative to central pictures. This means that, reasoning within the framework of the model, some information about the predicted timing of the gesture must have been available to the conceptualizer, enabling it to delay the speech more if the gesture is going to take longer to execute. The reverse also holds: the process responsible for planning the gesture must have had access to the amount of delay in speech. The error data show that the initiation of the movement and the duration of the forward motion *together* compensate for the delay of speech onset.

Second, the location of the peak syllable has a weak but systematic effect on the early timing of the gesture. This suggests that there is some form of information exchange between phonological encoding and pointing planning or even execution processes. However, both for the duration of the forward motion and the resulting apex onset time the only significant difference found is a difference between stress location 1 and later stress locations. Perhaps the location of the stressed syllable can only influence the timing of the pointing motion if the phonology of the word becomes available while the forward motion is still underway. Although the possibility of a direct motor-locking between pointing execution and articulation has been discarded in the analysis of results, it is not inconceivable that a more complex form of interaction between the motor systems of gesturing and articulation is responsible for the effect of stress location on the pointing. For instance, "beat" gestures, rhythmical up-and-down movements of the hand that do not carry meaning, have been shown to interact with peak syllables in intonational phrases. If a peak syllable co-occurs with a beat gesture, it is

[15]To investigate this hypothesis in more detail (e.g. to check whether in Experiment 2 the feedback can arrive in time to affect the hold duration) a model is needed in which specific assumptions about the timing of all relevant subprocesses are made.

[16]In the proposed architecture the claim is that only *obligatory* pointing gestures show such a synchronization pattern.

always with the *downward* motion of the beat gesture (McClave, 1994).

It is tempting to interpret these data, especially the error data from Experiment 2, as support for fully *interactive* models of speech and gesture production, as proposed by McNeill (1997). In interactive models, there is continual sharing of state information between gesture production processes and speech production processes. While these results certainly show that there are interactions between the planning of gesture and the planning of speech, there are a number of reasons why using a fully interactive approach is not necessarily going to be the best way to account for these results. First, "interactiveness" defines a *class* of models — not a specific model for gesture and speech. In other words, having a fully interactive model might allow for state information from any process to be shared by other processes, but it still does not specify what information is shared how and when. Second, the results from this study reveal different types of synchronization mechanisms. As the analysis in section 3.3.1 shows, the post-stroke hold, for instance, adapts more to the location of the peak syllable in the speech than earlier phases of the pointing gesture such as planning and execution do. In the modular approach, this can be explained in a natural way by assuming that feedback from the comprehension system is involved in "holding" the gesture. In fully interactive models one would have to explain why the post-stroke hold of a gesture is more adaptive to the timing of speech than the planning and execution stages of a gesture are.

HOW MAKING GESTURES HELPS YOU SPEAK

CHAPTER 4

4.1 Introduction

When people speak, their speech is often accompanied by spontaneous hand gestures. For some gestures, such as pointing gestures, or the "yay big" [17] gesture, it is clear that the gesture transmits essential information to the listener. On the other hand, so-called "beat gestures", rhythmical hand movements without any depicting properties, do not seem to add much information to the speech. Between these two extremes there are what McNeill (1992) calls *iconic* gestures. The defining property of an iconic gesture is that there is a shape-meaning relation — some aspect of the topic of the speech is represented in the shape of the gesture. For the remainder of this study, the word "gesture" is intended to mean "iconic gesture".

For iconic gestures, it is not straightforward to determine what their function actually is. In an extensive review of the literature, Kendon (1994) reaches the conclusion that gestures have a *communicative* func-

[17] This expression is used in American English in combination with a gesture to indicate a size. It roughly translates to "this big" with the difference that "this big" does not necessarily require an accompanying gesture, whereas "yay big" does. A typical example is the sentence: "The fish I caught was [gesture] yay big".

tion: they are produced for the benefit of the listener. Although this may seem a trivial conclusion, it is not undisputed. Krauss et al. (1991) and Rimé and Schiaratura (1991) have defended the view that gestures are not made for the benefit of the listener, but rather for speaker-internal reasons. According to Morrel-Samuels and Krauss (1992), (most) gestures are not communicative acts, but are performed by speakers in order to facilitate lexical retrieval (see also Krauss et al., 1996).

An argument often used by proponents of the non-communicative view is that people frequently gesture while they are on the telephone or, more generally, when there is no visual contact between speaker and listener. This phenomenon could be adequately explained by the theory that gestures facilitate the speaking process. If people gesture for speaker-internal reasons, they will do it on the telephone as well. But speaker-internal reasons cannot be the only reason for gesturing: speakers gesture more when they have visual contact with the listener (Cohen, 1977). Also, the fact that people gesture on the telephone does not rule out the possibility that gestures are communicatively intended. Gesturing could be so intricately linked to speaking that it is hard to suppress gesturing when speaking on the telephone. After all, speaking with someone whom you can't see is, from the viewpoint of human evolution, a very recent invention. It is conceivable that if gesturing is deeply integrated with the speaking process, the mere fact that the addressee is invisible is not sufficient to cause people to suppress gesturing.

Although there is disagreement in the literature about whether or not iconic gestures are communicatively intended, the issue whether making gestures facilitates the speaking process can be treated as an independent issue. Even if iconic gestures are, in fact, communicatively intended, it is still possible that they facilitate the speaking process.

An interesting question is how gesturing can facilitate the speaking process at all. According to Krauss et al. (1991) the motoric representation of a certain concept functions as a "cross modal prime" that helps to find a word in the mental lexicon. Indeed, speakers sometimes make repeated gestures when they are trying to find a word that seems to be inaccessible to them. But for the majority of gestures, there does not

seem to be a problem in finding the words that are produced at the time the gesture is made. Another problem with the proposal by Krauss et al. is that the meaning of gestures does not generally correspond with single words. McNeill (1992) presents data illustrating this point. For instance, in McNeill's corpus, someone says "He climbs up the drain spout of the building" while the "hand rises up with the first and second fingers wiggling, depicting the character's rising and clambering movement." (McNeill, 1992, p.106). This gesture contains far too much information to be encoded in a single lexical item. At best, it could have had the entire phrase as affiliate, but that would lead to another problem: gestures often reveal information that is not in the speech at all. To quote another example from McNeill: a participant talking about a Sylvester and Tweety cartoon explains one of the attempts of Sylvester to catch Tweetie, and says "the last one he tries is to ... is to walk across", while the participant makes an iconic gesture depicting the wires that were used to walk across on in the cartoon. However, the speaker did not mention the wires at that point of the conversation (McNeill, 1992, p.204). We must assume that the wires were somehow represented in the speaker's mind at the time the gesture was made. Therefore, it is likely that these gestures were generated from imagistic representations, and not from the linguistic representations underlying the speaking process.

If the assumption that gestures are generated from imagery is correct, then generating a gesture involves *accessing* the imagery. Accessing the imagery might in turn activate or re-activate the very same imagistic representation that needs to be inspected by the speech production system in order to generate speech. Aspects of the imagery must be translated into propositional format in order to be expressed in speech (cf. Levelt, 1989). If gesturing accesses and activates the imagistic representation, the inspection of the imagery by the speech production system is facilitated. This hypothesis will be called the *retrieval* hypothesis. Note that the hypothesis is not that gesturing facilitates the speaking process itself, but only enhances the accessibility of the imagistic information that the speech is about. In the words of Freedman (1977): "... the status of the motor act is like that of a catalyst: it evokes vague sensorimotor images, but leaves the verbal operations to more advanced

cognitive structures within the person".

Another way in which gesturing might help the process of speaking is through facilitating the process of encoding the imagistic information underlying the gesture into linguistic representations. Krauss et al. (1991) and also Krauss et al.'s (1996) claim that perceiving one's own gesture will activate features of a concept that is to be expressed, thereby cross-modally facilitating the retrieval of the proper lexical item. Another more recent proposal is by Kita (to appear). Kita hypothesizes that gesturing *reorganizes* imagistic information such that it is better suited for expression in speech — in other words, gesturing facilitates "thinking for speaking" (cf. Slobin, 1987). While different in many respects, the proposals by Krauss et al. and Kita share the idea that gesture helps the process of generating speech itself, and not just the retrieval of the imagistic information that speech is to be generated about. The hypothesis that gesturing helps the generation of speech itself is called the *encoding* hypothesis.

The most direct way of testing the encoding hypothesis would be to prevent participants from gesturing and look at the effects it has on their speech. This is precisely what a number of authors have done. Graham and Argyle (1975) presented geometrical line drawings to what they called "encoders". Encoders were either native speakers of Italian or native speakers of English. The task of the encoder was to describe those drawings to a "decoder" who had to reproduce the drawing. In one condition the encoder was allowed to gesture, while in the other he or she was not. The accuracy of the reproduction was higher when the encoder was allowed to gesture. This effect was even stronger for those drawings that were rated to be of low codability, demonstrating that the information presented in the encoder's gesture had a positive effect on the communication between encoder and decoder. No effects on the content of the speech were found. In Graham and Heywood (1975) essentially the same experiment was performed with only English speaking participants. They coded a large number of speech related dependent variables, of which only a few turned out to differ between the gesture and the no-gesture condition. Specifically, the elimination of gestures led to an increase in expressions describing spatial relations

and to a decrease in the number of demonstratives. Also, the time spent pausing (in speech) increased in the no-gesture condition. As the authors note, these findings need not be explained by the assumption that the production of speech is facilitated by gesturing. Rather, it is likely that the increased number of phrases describing spatial relations and the increased pausing time are a compensation for not being able to use the gesture modality, as is also suggested by Kendon (1972) and by De Ruiter (to appear).

There are, however, authors who have claimed to have found an effect of preventing gesturing on speech production itself. Rimé, Schiaratura, and Ghysselinckx (1984) let their participants engage in free conversation about certain predefined general topics. During the second half of the conversation, the head, hand and arm movements of the participant were immobilized by devices attached to the armchair of the participant. It was found that the vividness of the imagery in the speech *decreased* when the hands were immobilized. At first sight these results seem to contradict the aforementioned findings by Graham and Argyle (1975) and Graham and Heywood (1975): they found an *increase* in "spatial" speech, while Rimé et al. found a decrease in spatial speech. However, a crucial difference is that in the studies by Graham & Heywood and Graham & Argyle, the participants were requested to speak about the presented line drawings, while in the study by Rimé et al. participants were much more free to select the content of their speech. Assuming, again, that gesture is a communicative device that serves especially well to transmit spatial information, in the studies by Graham & Argyle and Graham & Heywood, participants were forced to compensate for the lack of gesture by producing more spatial descriptions in speech, while in the study by Rimé et al. participants could avoid talking about topics containing spatial information, thereby circumventing the problems the participants of Graham & Argyle and Graham & Heywood had.

Finally, Rauscher, Krauss, and Chen (1996) prevented their participants from gesturing as well. The participants in their study had to describe cartoon animations to listeners, while during half of the time they were not allowed to move their hands. Their findings were (1) that speech with spatial content was less fluent when gesturing was not permit-

ted, (2) speech without spatial content was not affected, and (3) that the frequency of (non-juncture) filled pauses in the speech increased in the no-gesture condition, but only when the participants were producing speech with spatial content. The authors conclude from these findings that gesture facilitates access to the mental lexicon, for the effects of preventing gesture are similar to those of word-finding difficulties. However, their results can easily be interpreted as evidence that gesture functions as a communicative device, as in the studies mentioned previously. Given that the gesture modality is much more efficient in expressing spatial information, the loss of fluency in the no-gesture condition is predictable: the generation of speech with spatial content needs to be adapted (i.e. be more accurate and elaborate) when the gesture modality is unavailable. If the content of the speech is not spatial, this problem does not occur, which is exactly what the authors found. The authors' conclusion that their findings indicate that gesturing facilitates lexical access therefore seems unwarranted.

These studies, in which the speech of participants was compared in a gesture and no-gesture condition, all have one important aspect in common: the speaker and listener could see each other during the experiment. Given that the results of these four studies can be elegantly and adequately explained from the assumption that gesture is used as a communicative device, the conclusion is that they do not reveal any facilitatory function of gesture on speech per se.

The following experiments are aimed at testing both the hypothesis that gesturing facilitates the retrieval of imagery (the *retrieval* hypothesis) and the hypothesis that gesture facilitates the encoding of imagery in linguistic format (the *encoding* hypothesis). Rather than preventing participants from gesturing and studying the effect on speech, in these experiments a reverse approach was taken: manipulating the conditions under which the speaker has to produce speech, and looking at the effect it has on the *frequency* of gesture.

In order to avoid the confounding effect of potential gestural communication, in the experiments reported below the speaker and listener were prevented from seeing one another.

4.2 Experiment 3

The aim of this experiment is to test both the retrieval hypothesis and the encoding hypothesis in a single experiment. For this purpose, a large number of spontaneous gestures were collected under experimental conditions. The participants were required to describe line drawings to another participant. The principal manipulation for testing the retrieval hypothesis was that the description of the line drawings was performed either from memory or from a screen. The encoding hypothesis was tested by having half of the line drawings to be easily describable, and the other half hard to describe.

The prediction of the encoding hypothesis is that the pictures that are hard to describe will be accompanied by more gestures than the pictures that are easy to describe. The retrieval hypothesis predicts that participants will make more gestures when they have to describe pictures from memory, than when they can see them on the screen.

4.2.1 Method

In order to collect a large number of spontaneous gestures, it was necessary that the participants were not aware of the nature of the experiment. Therefore, participants were presented with pictures on a computer screen which they had to describe in such a manner that another participant (who was a collaborator) behind a curtain could draw these pictures. The participants were told that they took part in a communication experiment.

The two principal manipulations in the experiment were (1) half of the pictures were "hard" to describe and the other half were "easy" to describe (see below), and (2) half of the pictures were described by the participants while they were visible on the computer screen, and the other half were described from memory.

Participants

22 native speakers of Dutch participated in the experiment. There were 4 male and 18 female participants. All participants were right handed except for one male participant who described himself as ambidextrous. The participants were students from the University of Nijmegen, and were paid for their services.

Procedure

Participants entered the experimental room in pairs. One of these participants was always the same collaborator, which was unknown to the other participant. The participants were introduced to each other, and one of them was assigned the role of "describer", and the other one the role of "drawer". The collaborator was always chosen as "drawer". The instruction for the describer was to sit down behind a computer screen and look at the presented pictures. In the SCREEN condition, (which was either in the first or the second half of the experiment, depending on the experimental group) the participant was requested to describe the picture on the screen to the drawer in such a way that she (the drawer) could reproduce the pictures using pencil and paper. In the MEMORY condition (the other half of the experiment) the describer was asked to first memorize the picture, then press a key to make the picture disappear, and only then start describing the picture to the drawer.

The drawer was requested to sit at a table and try to draw the pictures on paper from the description given by the describer. Both participants were told that if something was unclear, the drawer could ask questions. Before the experiment, however, the collaborator (who always played the role of the drawer) was trained during a pilot experiment to minimize feedback. At appropriate moments, the drawer produced an occasional "hm-hm", "yes" or "go on", but only if the describer said something truly incomprehensible the drawer would ask for clarification. The describer determined the pace of the experiment by pressing a key on the keyboard of the computer to see the next picture on the screen. There was a curtain between the drawer and the describer, so

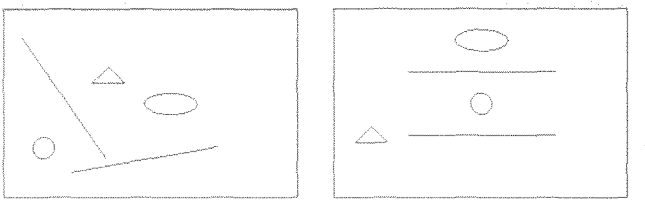
that they could hear, but not see each other. This was necessary to ensure that the gestures made by the describer were not explicitly or implicitly made for communicative purposes. Because there was no visual interaction between describer and drawer, the gestures produced by the describer were comparable to gestures made during telephone conversations.

The participants who were describing the pictures were videotaped using two cameras. One was mounted on the ceiling, straight above the describer, and the other was placed in front and at a slight angle to the right of the describer. Although no attempt was made to hide the video camera, none of the participants were aware that they were videotaped. After the experiment, the participants were informed about the fact that they had been recorded on video, and asked to sign a written agreement in which they gave the author permission to use the collected material for research purposes. None of the participants revealed any objection to having been recorded on video, and they all signed the agreement.

Materials and design

The pictures used in the experiment were simple line drawings, consisting of an ellipse, a circle, a triangle and two straight lines. The distinction "hard to describe" vs. "easy to describe" was realized by manipulating the relative placement of the five elements of the drawing. For the EASY pictures, the different elements were placed above, below, to the left, or to the right of the other elements. Furthermore, the two lines in the picture would always be either horizontally or vertically placed. For the HARD pictures, the five elements were placed in essentially random locations. For instance, the ellipse would be to the left of and below a line, but the line would be diagonally placed so that it was not possible to tell where the ellipse was just by using relative spatial predicates like "above" or "to the left of" (see Figure 4.1). In Appendix C, all pictures used in Experiment 3 are reproduced.

The other experimental manipulation was that pictures were either described directly from the screen (the SCREEN condition) or from mem-



(a) Hard (b) Easy

Figure 4.1: Examples of pictures used in Experiment 3

ory (the MEMORY condition).

The design was therefore 2×2, crossing both SCREEN/MEMORY and HARD/EASY. Each participant would receive all four conditions, but to counterbalance for potential presentation order effects, four experimental groups were created. In groups 1 and 2 two SCREEN sets were run first, and then two MEMORY sets. In groups 3 and 4 first the two MEMORY sets were presented, and then the two SCREEN sets. Furthermore, in groups 2 and 4, the order in which the pictures appeared was reversed. Each picture set consisted of three HARD and three EASY pictures. The sequence in which the HARD and EASY pictures occurred within a set were randomized, with the constraint that no three pictures of the same type would be presented in succession.

Analysis

The data from six participants who did not make any gestures during the experiment were excluded from the analysis, because their gesture behavior did not supply information to either verify or falsify the hypotheses under investigation.

For the other sixteen participants, a transcript was made of their speech, mainly for performing the word count, and the number of representational gestures were counted for each described picture. Representational gestures were defined to be hand movements that entertained

some relation to the elements of picture. This relation to the picture often needed to be established using the concurrent speech. That is, if a participant would say "the triangle is to the right of the square" while putting out the right hand to the right of her shoulder, this hand motion was interpreted to be an iconic gesture. If the same movement would have been made (for instance to chase away a fly, or to express despair) while saying "this picture is difficult to describe", the hand movement was not considered to be an iconic gesture. In case of repetitive gestures, for instance drawing an ellipse in the air several times in a row without the hand retreating to a resting position in between, the gesture was counted once, unless the accompanying speech was different for each repetition. As an example, if someone said "There also was an ellipse" accompanied by three repetitions of an ellipse drawing gesture, it was counted as one gesture. However, if someone said "There also was an ellipse [gesture] and that ellipse [gesture] eh, I mean, the ellipse [gesture] ..." with each gesture being an ellipse drawing gesture, it was counted as three gestures.

For the word count, interjections such as "eh" or "ehm" were not counted as words.

Results

The main dependent variable in this experiment is the rate of gesture, that is the number of gestures per word made by the participants during his or her description of the pictures. This unit was chosen because the alternative measure, gestures per second, would be influenced by the speech rate: if participants slow down their speech the gesture rate would go up. The gestures per word measure does not have that problem.

First, it was important to check whether the experimental manipulation of the HARD/EASY dimension was effective. The mean speech rate in words per second for the HARD pictures (1.75 w/s) was significantly lower than the speech rate for the EASY pictures (1.88 w/s, $t_1(15) = 4.64$, $p < .001$, $t_2(10) = 4.70$, $p < .001$, one tailed), showing

that this manipulation was, indeed, effective.

The gesture rates in the four conditions are shown in Table 4.1.

	EASY	HARD
SCREEN	3.5	4.0
MEMORY	4.4	4.4

Table 4.1: Gestures per 100 words in Experiment 3

The main effect of EASY/HARD was nonsignificant ($t_1(15) = -.93$, $p = .184$, $t_2(10) = -.74$, $p = .24$, one tailed)[18]. The gesture rate did not increase when describing the pictures that were hard to describe.

However, the main effect of SCREEN vs. MEMORY was significant by participant, and marginally significant by item ($t_1(15) = -1.99$, $p = .03$, $t_2(11) = -1.70$, $p = .058$, one tailed). The interaction between the factors EASY/HARD and SCREEN/MEMORY was nonsignificant (both F_1 and $F_2 < 1$), but the difference in gesture rate between HARD and EASY pictures within the SCREEN condition is approaching significance in the analysis by participant ($t_1(15) = -1.66$, $p = .059$, $t_2(10) = -1.07$, $p = 0.16$, one tailed).

4.2.2 Discussion

These results support the hypothesis that gesture helps in retrieving spatial information from memory: talking from memory leads to a higher gesture rate. However, interpreting the nonsignificant difference between the HARD and EASY pictures is complicated by the fact that there is a marginally significant ($p = .065$) simple effect of HARD/EASY within the SCREEN group. First, the null effect for EASY vs. HARD pictures could be nonsignificant because the power of the experiment was too low. Second, it is possible that the difference

[18]The item analysis was performed with pictures as unit of analysis.

between the hard and easy pictures was not large enough to detect an effect on the gesture rate.

Therefore, a second experiment was performed. The second experiment used only the SCREEN condition of the first experiment, because that was the condition in which the trend in gesture frequency was observed in Experiment 3. The EASY pictures were made easier, and the HARD pictures harder. Also, the statistical power of this second experiment was increased by using more participants and pictures per condition.

4.3 Experiment 4

Experiment 4 is essentially a replication of Experiment 3, the main difference being that the memory condition was removed: the participants only described pictures directly from the screen. The aim of this experiment is to see whether the trend for the HARD/EASY factor in Experiment 3 is a reliable effect. The encoding hypothesis would predict such an effect, while the retrieval hypothesis would not.

4.3.1 Method

Participants

23 native speakers of Dutch participated in the experiment. None of them had participated in Experiment 3. Six participants were male, and seventeen were female. One male participant was left handed, all the others were right handed.

Procedure

The experimental procedure was identical to the one used in Experiment 3. The same collaborator was used to draw the pictures. The

only difference was that participants only did the SCREEN condition of Experiment 3, with more and different pictures.

Materials and design

In order to enhance the difference between the EASY and the HARD pictures, the three easiest and the three hardest pictures from Experiment 3 were used. In addition, two even more easy pictures were added to the EASY group of pictures, and two even harder pictures were added to the HARD group of pictures. To determine what pictures were "easiest" and "hardest" in Experiment 3, the average speech durations of the picture descriptions were calculated, the assumption being that the ease of description for pictures was reflected in the average speaking time needed to describe them. The three easiest pictures from Experiment 3 were c1, c6, and c7 (see Figure C.1). They appear to be easy because three or four elements of these pictures lay in a straight line. Therefore, the two new pictures for the EASY group (see Figure C.3) were made such that all the elements are either horizontally or vertically organized in a straight line. For the hardest pictures, n6, n7, and n8, the hardness of these pictures appeared to be that there were no clear groupings of the elements. Therefore, the two new pictures were attempted to be made such that there were even less visual groupings present (see Figure C.4).

The order of presentation of the pictures was randomized such that no three hard or easy pictures would appear consecutively. Because there was only one factor in this experiment, two experimental groups were created. In the second experimental group, the order in which the pictures had to be described was reversed with respect to the first group.

Analysis

Of the original 23 participants, the results of nine were not analyzed. One participant made finger drawings on the table in front of her. Since it was unclear whether to interpret these finger drawings as gestures, the data from this participant were excluded from the analysis. Eight

other participants did not make any representational gestures during the descriptions, so they were excluded from the analysis as well.

The pre-analysis of the data was otherwise performed in exactly the same way as in Experiment 3.

Results

As in the previous experiment, the speech rate of the participants was investigated to see whether the experimental manipulation was indeed effective. The average speech rate for the EASY pictures was 2 words per second, while for the HARD pictures it was 1.86 words per second ($t_1(13) = 2.70$, p = .009, $t_2(8) = 2.62$, p = .015, one tailed). Also, the difference in speech rate was larger than in the previous experiment, indicating that the difference in ease of description was larger too.

The gesture rates for the EASY and HARD pictures in gestures per word are 0.055 for the EASY pictures, and 0.053 for the HARD pictures. This difference is in a direction opposite to the one predicted by the encoding hypothesis. The difference is also nonsignificant ($t_1(13) = 0.27$, p = .8, $t_2(8) = 0.55$, p = .6, two tailed).

4.3.2 Discussion

Concluding, even with more pictures, and a larger difference between the HARD and EASY pictures, the HARD/EASY manipulation did not result in a difference in the gesture rate. It would still be possible to maintain that gesturing helps in encoding the imagistic information into linguistic format, but only if one assumes that encoding the EASY pictures into linguistic (propositional) format would be as hard as encoding the HARD pictures. I believe this to be unlikely. The average time it took the participants to complete a description of the EASY pictures in this experiment was 73.5 seconds, while for the HARD pictures it was 148.5 seconds, which is twice as long. Note that the number of elements in the pictures was the same. The difficulty in describing the

HARD pictures was to explain to the drawer where the elements had to be placed. If gesturing were helpul in the encoding of imagistic information into propositional format, this persistent null effect on the gesture rate would be unlikely to occur.

4.4 Conclusions

The null effect with regard to the difficulty of the pictures is hard to interpret. All null effects have to be interpreted cautiously, but in this case there is the additional problem that even the EASY pictures in this study could still have been very hard to describe for the participants. In that case, all pictures in both experiments would have been HARD, and therefore the data would show no visible differences in gesture rate.

However, we can conclude that these results did not provide any evidence for the encoding hypothesis, which stated that gesturing facilitates the process of verbalizing spatial information. The difference in ease of description between the hard and easy pictures was quite large, as revealed by the lower speech rate with the hard pictures. If gesturing helps to translate imagistic information into linguistic representations, it is surprising to find that the hard/easy manipulation in the experiments did not have *any* effect on the gesture rate.

The finding that people gesture more when they describe something from memory, even when the listener cannot see them, suggests that gesturing helps to retrieve images from memory. Facilitating the retrieval of spatial information will also be beneficial to the speaking process as a whole. I therefore conclude that facilitating access to spatial memory is one of gesture's functions.

Further research is needed to investigate the facilitatory function(s) of gesture in more depth. For example, participants could be prevented from gesturing under the same conditions as in Experiment 3 in this study. If, in such an experiment, the fluency of spatial descriptions in speech would decrease as a result of preventing gesture, this would provide strong evidence for a possible facilitatory function of gesture.

However, in the "preventing gesture" paradigm it will be very difficult to distinguish possible effects of the encoding hypothesis from possible effects of the retrieval hypothesis. If true, both hypotheses would predict a detrimental effect on the fluency of speech, but to know exactly what effect causes what kind of disfluency, very detailed assumptions have to be made about gesture and speech production and their interactions, both with each other and with the retrieval of imagery.

CONCLUSIONS

CHAPTER 5

In the previous three chapters, the phenomenon of speech-related gesture was investigated both theoretically and experimentally. In this chapter, I will discuss the implications of the findings, relate them to one another, and suggest possibilities for further research.

In the first chapter, an information processing architecture was formulated for the simultaneous production of speech and gesture, called the Sketch Model. The main advantage of formulating a model is that it can be used to generate more detailed predictions than "verbal" theories usually allow for. Besides generating predictions, the model serves as a "map" of what parts of the theoretical territory have been left uncovered by processing theories. This "map" also raises some new issues concerning processes and representations.

For instance, it is known that certain classes of gestures are subject to a degree of conventionalization, like for instance the handshape used in pointing gestures (Wilkins, 1997). These conventions are shared within a language community, which implies that somehow this shared knowledge must be represented in the speaker's mind. In the Sketch Model it is assumed that a special knowledge store, called the gestuary, contains the information that is shared amongst speakers of the same language. Pointing gestures, however, differ from emblems in that they exhibit not only conventionalized properties but also *analog* properties: what one points at is determined by one's communicative intention and the

physical environment. Therefore, there has to be a process that *merges* shared knowledge and locally determined information (in this example, the target of the pointing) in one gesture. In the Sketch Model, it is therefore assumed that the gestuary contains gestural *templates* that specify a number of parameters for a certain (conventional) gesture, while leaving other parameters free. The other parameters are to be filled in when the gesture is actually performed. A similar mechanism is proposed for the merging of different sources of information into one gesture. A possible direction for further research would be to investigate if the "degree of freedom" approach proposed in the Sketch Model is compatible with existing theories about motor programming.

The Sketch Model is an extension of Levelt's (1989) model for speaking. Because the modularity assumption of the latter model is largely adopted, the Sketch Model is vulnerable to falsification. Indeed, the prediction that non-obligatory gestures are only synchronized at an early level of gesture/speech planning, was shown to be false in chapter 3. In that chapter, pointing gestures and their accompanying vocalizations were recorded in an experimental design that aimed at testing Kendon's (1980) claim that gestures are synchronized with the phonological peak syllable of the accompanying speech. While the temporal location of the lexically stressed syllable did not have any effect on the gesture timing, the location of the intonational peak syllable did. Another important finding from chapter 3 is that the timing of the pointing had an effect on the onset of the speech fragment: if the pointing was to a more distant location (causing the gesture to be longer in duration) the speech started later.

The Sketch Model needs to be modified to accommodate these results. The adaptation from speech to gesture (speech onset is delayed if the gesture duration will be longer) implies that the process responsible for initiating a speech fragment takes into account the amount of time needed to execute the gesture. It seems therefore necessary to assume that the gesture planner, after having constructed a motor program for the gesture, sends a "resume" signal to the conceptualizer. Upon receiving this signal, the conceptualizer can send the preverbal message to the formulator. See Figure 5.1 for an overview of the revised model.

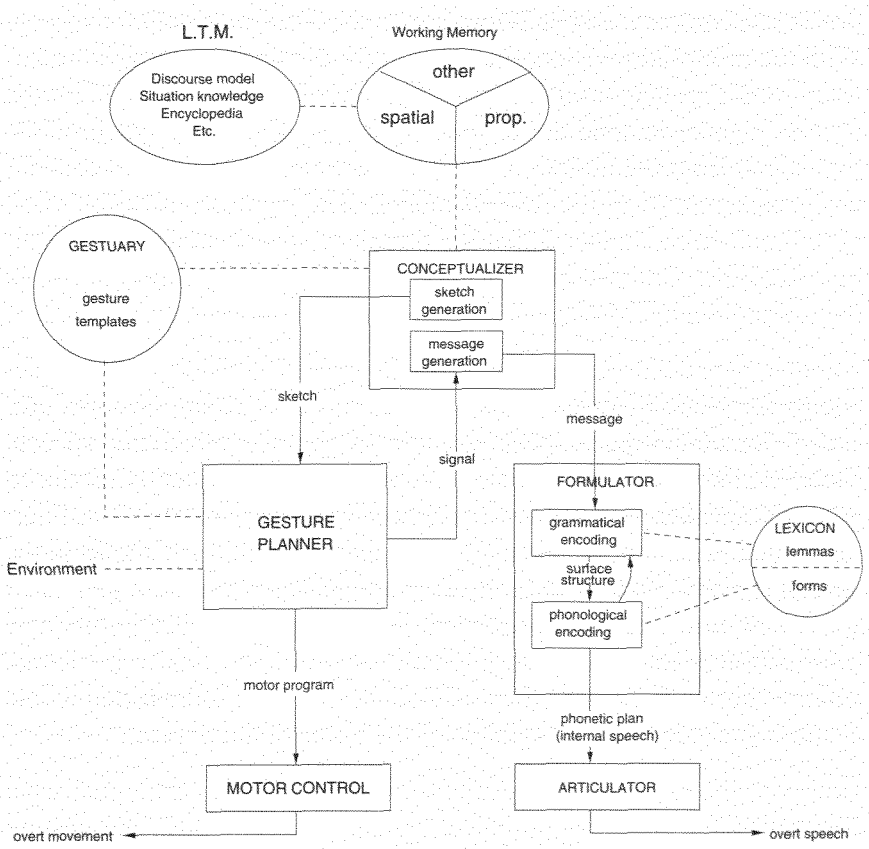


Figure 5.1: The revised Sketch Model

This revision of the Sketch Model would automatically lead to a more principled account of the phenomenon that gesture always precedes its accompanying speech. In the original Sketch Model the assumption is that gesture and speech are initiated simultaneously, and that gestures are executed earlier because the processing involved in gesture production is less complex, and therefore less time consuming, than for speech. With the proposed revision, the account for the finding that gesture onset precedes speech onset is simply that the conceptualizer waits for the gesture planner to send the resume signal. Note that the synchronization signal from the motor control process to the phonological encoder has been removed in the revised model. The result from Levelt et al. (1985) and from chapter 3 that the timing of speech adapts to the timing of gesture can also be explained by the new signal mechanism in the revised Sketch Model.

Another beneficial side effect of the suggested revision is that it accounts for Nobe's (1996) finding that the onset of iconic gestures preceded the phonological peak syllable of the accompanying speech. McNeill (1992) has shown that iconic gestures are frequently produced when the speaker presents new (as opposed to given) information. Fragments that present new information are marked with pitch accent in the speech (Levelt, 1989, p.151). If (a) iconic gestures are often generated when new information is presented, and (b) speech that presents new information is marked by an intonational peak, and finally (c) iconic gestures normally precede their affiliate in speech, it follows that iconic gestures precede the peak syllable of the intonational phrase.

However, the experimental results from chapter 3 also show that what was called the *strict* phonological synchrony rule holds: the gesture not only precedes, but also *covaries* with the peak syllable. This finding cannot be incorporated at the level of the conceptualizer or gesture planner, because these processes do not have access to information at the word or syllable level. As mentioned in chapter 3, these small but significant adaptations of gesture timing to the timing of the peak syllable could be the result of low level interdependencies between the respective motor control processes for articulation and limb movement. An interesting question for future research is whether and if so,

how these motor interdependencies operate. One possibility that has not been investigated yet is that breathing is an underlying factor in the strict phonological synchrony rule. Raßler, Ebert, Waurick, and Junghans (1996) have used finger tracking (a process very similar to pointing) to demonstrate that there is a two-way interaction between the phase of breathing and the accuracy and velocity of the hand movement. The authors also showed that the velocity and accuracy of the hand are maximal during the middle period of a respiratory half-cycle. When peak syllables are produced in speech, the outgoing airflow is maximal. The motor system could therefore plan the moment of maximal exhalation to co-occur with the outward motion of the hand. Since the moment of maximal exhalation varies with the location of the peak syllable, the phonological synchrony rule might be explained by a phase locking mechanism at the level of lower level motor planning, instead of a higher level synchronization process.

The most intriguing result from chapter 3 concerns the hesitations and interruptions in the productions of the experimental participants. In case of a speech error, gesture has been shown to restore its timing relation with speech. The delayed onset of the speech was compensated for by a combination of a delayed onset *and* a slower execution of the gesture. This adaptation resulted in a temporal difference between the onset of the apex and speech onset that was almost identical to the one in normal trials. The duration of the interruption or hesitation itself was compensated for by holding the pointing hand out longer. This lengthening of the hold phase of the gesture in case of an interruption of the speech flow supports an important assumption in the original Sketch Model (and its revised version): the gesture is "held" until the conceptualizer receives the feedback that the accompanying speech is produced successfully. However, the adaptation of both gesture onset and execution speed, to adapt to the delay in speech onset, is more difficult to accommodate, both in the original Sketch Model and its revised version. Somehow, during the simultaneous planning of gesture and speech in the conceptualizer, the gesture is not only delayed, but also *slowed down* when there are problems in speech production. As mentioned in chapter 3, especially this last finding could lead one to believe that gesture and speech are fully interactive (McNeill, 1997). How-

ever, full interactiveness would be computationally both complex and expensive. It would imply that every few milliseconds, the progress of the gesture production would have to be compared with the progress of the speech production, and adapted accordingly. Not only would it be very difficult to formulate a model that does this, it would also lead to far more computational effort than is needed to obtain synchronization. For the adaptation phenomenon under discussion, it would be sufficient for the gesture planner to have information about the approximate delay in speech onset, so it could adapt the generated motor program accordingly. Further research, e.g. using a paradigm in which a larger number of speech errors are evoked on purpose, could perhaps reveal more detailed information about this type of temporal adaptation.

Finally, chapter 4 addresses the question why people produce iconic gestures. After an extensive review of the psychological literature, the conclusion was drawn that the available evidence indicates that people do this for communicative purposes (cf. Kendon, 1994). Although some authors have claimed that gestures are made for speaker-internal reasons only (Krauss et al., 1991; Rimé & Schiaratura, 1991; Rauscher et al., 1996), the results they present to support their view can all be explained adequately by assuming that iconic gestures are produced for communicative reasons. However, that does not exclude the possibility that gesture also has a facilitatory effect on the process of speaking.

In order to investigate the facilitatory effect of gesture on speech, without the possible confound of gestural communication, participants in the experiments from chapter 4 had to describe pictures to a listener they could not see. Two hypotheses were investigated. The first was that making gestures facilitates access to imagistic representations in the speaker's mind. This was called the *retrieval* hypothesis. The second hypothesis, called the *encoding* hypothesis, was that gesturing facilitates the process of speech generation itself. The result that people gestured more frequently when they had to describe pictures from memory supports the retrieval hypothesis. However, no support at all was found for the encoding hypothesis: the pictures that were very hard to describe evoked the same number of gestures (per word) as the pictures that were easy to describe. It is rather striking that such a huge differ-

ence in the "difficulty" of speech production did not have any effect on the gesture frequency. Although one has to be careful in interpreting a null effect, this result casts some doubt on both Krauss et al.'s (1996) idea that gesture facilitates lexical access, and Kita's (to appear) idea that gesturing facilitates the reordering of information for expression in speech.

The findings from chapter 4 are also in agreement with the Sketch Model. In the model, interactions between gesture and speech only take place at the level of the conceptualizer and not "below" that level. This means that the *retrieval* hypothesis is compatible with the Sketch Model, whereas the *encoding* hypothesis is not.

The studies reported in this dissertation reveal how complex the phenomenon of gesture is, and how little is known about it. However, the consistent application of modeling within the information processing framework supported by experimental testing has borne some fruit. Given enough experimental testing, all models can eventually be proven wrong, but rejecting a model is often more informative than confirming a theory.

SUMMARY

CHAPTER 6

During speaking, people often gesture. These gestures appear to be closely linked to the process of speaking. First, people gesture almost exclusively during speaking (and not, for example, during listening). Second, the meaning of gestures, if it can be identified, is directly related to the speech. The subject of this thesis is the production of gesture and speech, and the relation between them.

The gestures of the speaker are different from the gestures deaf people employ to communicate with each other. The sign language of the deaf is a real language. Like Chinese, or English, it has syntactic, semantic, and morphological rules that are shared by the speakers of that language. For most gestures made by speakers, these rules do not exist.

Gestures can be classified in many ways. The typology used in this thesis is by McNeill (1992). This typology distinguishes a number of categories. *Deictic* gestures are (pointing) gestures that indicate a certain location or direction. In Dutch or English the index finger is often used for deictic gestures, but in some cultures other fingers, the head, or the lips are used to make deictic gesture. *Emblems* are gestures that have a meaning that is shared by the speakers of a language. The thumb-up gesture to indicate that something is "OK", or the finger to the lip gesture that means "be silent" are examples of emblems. Another category of gestures is the *beat*. Beats are rhythmic up and down motions of the hand that appear to have no meaning. It is suspected that the rhythm

of these gestures is related to the phonology of the concurrent speech, but there is no convincing evidence to support that theory. Finally, an important category of gestures is formed by *iconic* gestures. The shape of these gestures has a meaningful relation to the subject of the speech. Making a spiraling motion with the index finger when talking about a vortex is an example of an iconic gesture.

The relation between gesture and speech is often investigated by carefully studying video recordings of speakers. This method has revealed many ways in which gesture and speech are related. The disadvantage of this methodology is that it is almost impossible to make general claims about the nature of gesture and speech, the reason being that gestures are highly variable. Different speakers make different gestures, but even within a single speaker talking about the same subject gestures vary considerably. Another problem is that many gestures are produced spontaneously and without conscious awareness. It therefore does not make sense to ask people to produce these gestures "on command". This methodological problem occurs mainly in the study of iconic gestures and beats. Because deictic gestures and emblems have a predictable shape, it is possible to use them in controlled experiments.

Psychological research into gesture and speech mainly focuses on the questions *why*, *how*, and *when* gestures are made.

In chapter 2 the question *how* people gesture is approached by formulating a blueprint for an information processing model. One of the advantages of formulating a model is that it results in an overview of the processing involved in the production of gesture and speech. The formulated model, the "Sketch Model", is an extension and adaptation of Levelt's model for speaking. In the Sketch Model the assumption is that gesture has a communicative function, just like speech. During the preparation of the speech the accompanying gesture is prepared as well. After the preparation phase, gesture and speech are processed largely independently of each other. This assumption differs from the one in Krauss et al. (1996), who assume that gestures are an *epiphenomenon* of speech. According to Krauss et al. gestures are not produced to add information to the speech, but to facilitate the process of speaking itself. This different assumption about the function of gesture results

in a different model than the Sketch Model. McNeill's (1992) *growth point* theory shares the assumption made in the Sketch model that gesture and speech cooperate in order to communicate. However, since growth point theory is not specified in terms of information processing, it is very difficult to make testable predictions. In contrast, the Sketch Model does make a number of clear predictions, especially about *when* people gesture, relative to the concurrent speech.

In chapter 3, Kendon's (1980) claim that gestures are produced simultaneously with the peak syllable in the intonational phrase is investigated. In the experiments, participants were requested to describe pictures and point at them. By recording both the speech and the motion trajectory of the pointing hand the temporal relation between gesture and speech could be investigated in detail. The first experiment revealed that the gesture was not affected at all by the temporal location of the stressed syllable in the speech. Whether participants said "de CAmera", [ENG: *the camera*] or "de krokoDIL", [ENG: *the crocodile*] did not influence the timing of the gesture. When in three-word utterances the focus of the speech fragment was varied from one word to the next, the resulting pitch accent placements did affect the timing of the gesture: the later the accented syllable was produced, the slower and longer the gestures became. Analyzing a number of speech errors resulted in an even more marked adaptation effect. If the speech was interrupted, the gesture adapted almost immediately, resulting in a temporal relation between gesture and speech that was identical to that relation in speech without errors.

Although a number of findings from this study support the Sketch Model, the results show that the interaction between gesture and speech is much tighter than is assumed in that model. Therefore the model is adapted in chapter 5. In the revised Sketch Model the assumption is that the generation of speech can only start after the process responsible for the planning of a gesture (the *gesture planner* in the model; see figure 5.1) has finished and sends a signal to the process that initiates the generation of speech (the *conceptualizer*). This adaptation of the model also explains in a simple way the phenomenon that gestures usually precede the affiliated speech by a small amount of time.

In chapter 4, the issue *why* people gesture is investigated. A number of authors claims that iconic gestures are not intended communicatively, but serve only to facilitate speaking. This could also be the reason why people gesture on the telephone, even though the listener cannot see their gestures. However, the experimental results that are presented mainly by Krauss and his colleagues to support their claims can all be explained with the straightforward assumption that gestures have a communicative function. Furthermore, the question whether gestures facilitate the speaking process is independent of the question whether or not they are intended communicatively. In the experiments of chapter 4 two hypotheses about the potential facilitative function gesturing has on speech are investigated. The first is that gesturing facilitates access to the representations in memory that the gesture is generated from. This is called the *retrieval* hypothesis. The second hypothesis is that gesturing facilitates the process of speaking itself, for instance because gesturing helps retrieving the correct words or concepts. This last hypothesis is called the *encoding* hypothesis. To test these two hypotheses, the participants were requested to describe pictures containing a number of geometrical figures in such a way that another participant could draw these pictures. The describer and the drawer could not see each other in these experiments, to prevent gestures from being made for communicative purposes. With half of the pictures, the participants had to memorize the pictures before describing them. The other half of the pictures was described directly from the screen. The pictures themselves consisted of one group that was easy to describe, and another group that was very hard to describe. The latter group of pictures was hard to describe because it was very difficult to say where the sub-figures of the pictures were located relative to one another. The result of this experiment was that participants made more gestures when they had to memorize the pictures than when they could see them. This result supports the retrieval hypothesis. However, there was no difference in the frequency of gesture between the hard and easy pictures, not even in a follow-up experiment that was devised especially to detect a possible effect of the difficulty of the pictures. Although it is dangerous to draw far-reaching conclusions from a null effect, this finding nevertheless seems to contradict the encoding hypothesis. If the encoding hypothesis is right, it

is unlikely that the large difference in difficulty between the the two groups of pictures has no effect at all on the frequency of gesture.

The findings from the experiments presented in chapter 4 are in agreement with the Sketch Model. In that model, the assumption is that interactions between gesture and speech are only possible in an early stage (in the *conceptualizer*). The retrieval hypothesis is therefore in agreement with the Sketch Model. The encoding hypothesis is not, because it assumes that gesture facilitates speaking at the level of the *formulator*, which is not possible in the Sketch Model.

Summarizing briefly the main findings of this thesis, the first result is that speech and gesture are tightly coupled. The timing of the speech influences the timing of the gesture, and the other way around. The intonation of the speech also plays a role in the timing of the gesture: Kendon's (1980) claim that gestures are synchronized with the intonational peak syllable in the speech has been confirmed. Another finding is that gesturing can facilitate speaking by (re)activating visual representations in short term memory. Finally, the production of gesture and speech, and their interaction, are such a complex phenomenon that using an information processing model is essential for making testable predictions.

BIBLIOGRAPHY

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review, 96*(1), 168–174.

Butterworth, B. L., & Beattie, G. W. (1978). Gesture and silence as indicators of planning in speech. In R. N. C. . P. T. Smith (Ed.), *Recent advances in the psychology of language 4: Formal and experimental approaches* (pp. 347–360). London: Plenum.

Calbris, G. (1990). *The semiotics of French gesture*. Bloomington: University of Indiana Press.

Churchland, P. (1986). *Neurophilosophy*. Cambridge, Massachusetts: MIT Press.

Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication, 27*, 54–63.

De Ruiter, J. P. A. (1995a). *Classifying Gestures by Function and Content*. Paper presented at 1995 Albuquerque Gesture Conference.

De Ruiter, J. P. A. (1995b). Why do people gesture at the telephone? In M. Biemans & M. Woutersen (Eds.), *Proceedings of the CLS opening Academic Year 1995-1996*. University of Nijmegen.

De Ruiter, J. P. A. (to appear). *The production of gesture and speech*. In McNeill (ed): Language and Gesture: Window into Thought and Action.

Efron, D. (1941). *Gesture and environment.* Morningside Heights, NY: King's Crown Press.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavioral categories. *Semiotica, 1,* 49–98.

Fast, J. (1971). *Body Language*: Pan Books, London.

Feyereisen, P. (1997). The competition between gesture and speech production in dual task paradigms. *Journal of Memory and Language, 36*(1), 13–33.

Fodor, J. A. (1983). *The Modularity of Mind.* Cambridge, Massachusetts: The MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Freedman, N. (1977). *Hands, words and mind: on the structuralization of body movements during discourse and the capacity for verbal representation* (chap. 2, pp. 109–132). New York, London: Plenum Press.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International journal of psychology, 10*(1), 57–67.

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of social Psychology, 5*(2), 189–195.

Hagoort, P., Brown, C. M., & Swaab, T. M. (in press). Lexical-semantic event-related potential effects in left hemisphere patients with aphasia and right hemisphere patients without aphasia. *Brain.*

Haviland, J. B. (1993). Anchoring, iconicity, and orientation in guugu yimithirr pointing gestures. *Journal of Linguistic Anthropology, 3,* 3–45.

Kendon, A. (1972). Some relationships between body motion and speech. In A. W. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (chap. 9). New York: Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–228). The Hague: Mouton.

Kendon, A. (1994). Do gestures communicate? a review. *Research on Language and Social Interaction, 27*(3).

Kita, S. (1990). *The temporal relationship between gesture and speech: A study of Japanese-English bilinguals.* Master's thesis, Department of Psychology, University of Chicago.

Kita, S. (1993). *Language and thought interface: A study of spontaneous gestures and Japanese mimetics.* Ph.D. thesis, The university of Chicago.

Kita, S. (1997). Two-dimensional semantic analysis of japanese mimetics. *Linguistics, 35*, 379–415.

Kita, S. (to appear). *How representational gestures help speaking.* In McNeill (ed): Gesture: an emerging field of study.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (vol. 28, pp. 389–450). Tampa: Academic Press.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61*(5), 743–754.

Levelt, W. J. (1989). *Speaking.* Cambridge, Massachusetts: MIT press.

Levelt, W. J., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24*, 133–164.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review, 98*(1), 122–142.

Levinson, S. C. (1996). *The body in space: cultural differences in the use of body-schema for spatial thinking and gesture.* Paper circulated for the Fyssen Colloquium: Culture and the uses of the body.

Marquardt, C., & Mai, N. (1994). A computational procedure for movement analysis in handwriting. *Journal of Neuroscience Methods.*

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research, 23*(1).

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*(350-371).

McNeill, D. (1987). *Psycholinguistics: a new approach*: Harper & Row.

McNeill, D. (1992). *Hand and Mind.* Chicago, London: The Chicago University Press.

McNeill, D. (1997). Growth points cross-linguistically. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 190–212). Cambridge University Press.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology; Learning, Memory and Cognition, 18*(3), 615–622.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading* (pp. 264–336). Lawrence Erlbaum & Associates.

Nobe, S. (1996). *Cognitive Rhythms, Gestures, and Acoustic Aspects of Speech.* Ph.D. thesis, Department of Psychology (Cognition and Communication), University of Chicago.

Pylyshyn, Z. W. (1979). Complexity and the study of human and artificial intelligence. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence* (chap. 2). Brighton, UK: The Harvester Press.

Raßler, B., Ebert, D., Waurick, S., & Junghans, R. (1996). Coordination between breathing and finger tracking in man. *Journal of Motor Behavior, 28*(1), 48–56.

Rauscher, F. B., Krauss, R. M., & Chen, Y. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226–231.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In R. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 239–281). Cambridge, England: Cambridge University Press.

Rimé, B., Schiaratura, L., & Ghysselinckx, A. (1984). Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. *Motivation and Emotion, 8*(4), 311–325.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*, 107–142.

Schmidt, R. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82*, 225–260.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana, Illinois: University of Illinois Press.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for invesetigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247.

Slobin, D. (1987). Thinking for speaking. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the 13th annual meeting of the Berkeley Linguistics Society meeting* (pp. 435–445.

Wheeldon, L., & Levelt, W. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language, 34*, 311–334.

Wilkins, D. P. (1995). *What's 'The Point'?: The significance of gestures of orientation in Arrernte.* Paper presented at the Institute for Aboriginal Development, Alice Springs, Australia.

Wilkins, D. P. (1997). *Why pointing with the index finger is not a universal (in socio-cultural and semiotic terms).* Paper presented at Max-Planck Workshop on Pointing Gestures.

APPENDICES

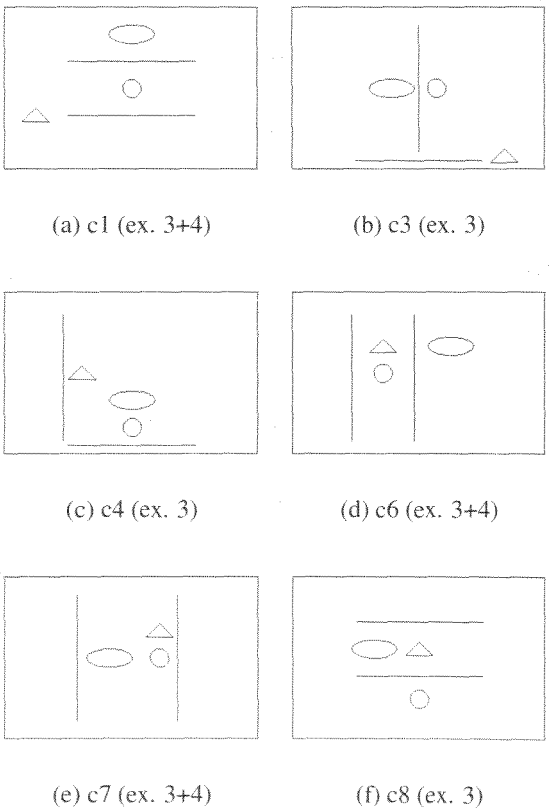
A Stimuli used in Experiment 1

Stimulus	stress	# syll.	gender	English
fakkel	initial	2	DE	torch
tijger	initial	2	DE	tiger
kuiken	initial	2	HET	chick
masker	initial	2	HET	mask
trompet	final	2	DE	trumpet
raket	final	2	DE	rocket
penseel	final	2	HET	brush
pistool	final	2	HET	gun
boterham	initial	3	DE	slice of bread
camera	initial	3	DE	camera
podium	initial	3	HET	stage
stadion	initial	3	HET	stadium
hagedis	final	3	DE	lizard
krokodil	final	3	DE	crocodile
etiket	final	3	HET	label
ledikant	final	3	HET	bed

B The picture groups of Experiment 2

Number	Far Left	Near Left	Near Right	Far Right
1	fakkel	penseel	podium	hagedis
2	podium	fakkel	hagedis	penseel
3	penseel	hagedis	fakkel	podium
4	hagedis	podium	penseel	fakkel
5	masker	raket	boterham	ledikant
6	boterham	masker	ledikant	raket
7	raket	ledikant	masker	boterham
8	ledikant	boterham	raket	masker
9	tijger	pistool	stadion	krokodil
10	stadion	tijger	krokodil	pistool
11	pistool	krokodil	tijger	stadion
12	krokodil	stadion	pistool	tijger
13	kuiken	trompet	camera	etiket
14	camera	kuiken	etiket	trompet
15	trompet	etiket	kuiken	camera
16	etiket	camera	trompet	kuiken

C Stimuli used in Experiment 3 and 4

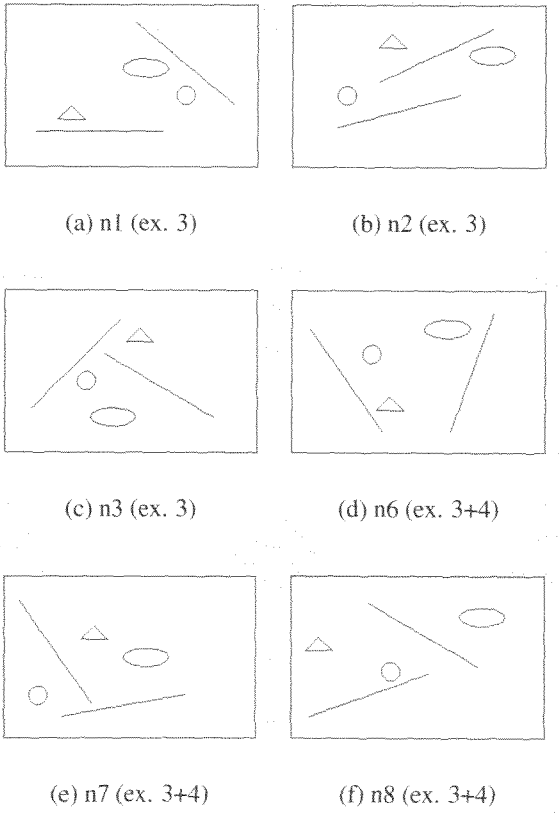


(a) c1 (ex. 3+4) (b) c3 (ex. 3)

(c) c4 (ex. 3) (d) c6 (ex. 3+4)

(e) c7 (ex. 3+4) (f) c8 (ex. 3)

Figure C.1: The EASY pictures used in Experiment 3

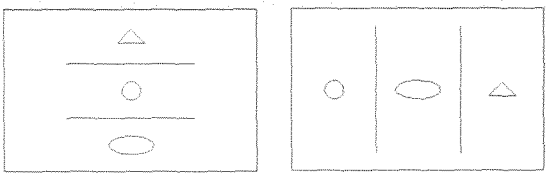


(a) n1 (ex. 3) (b) n2 (ex. 3)

(c) n3 (ex. 3) (d) n6 (ex. 3+4)

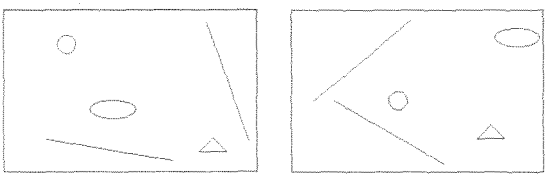
(e) n7 (ex. 3+4) (f) n8 (ex. 3+4)

Figure C.2: The HARD pictures used in Experiment 3



(a) ca (b) cb

Figure C.3: The extra EASY pictures used in Experiment 4



(a) na (b) nb

Figure C.4: The extra HARD pictures used in Experiment 4

GEBAAR- EN SPRAAKPRODUCTIE

Samenvatting

Tijdens het spreken maken mensen vaak gebaren. Deze gebaren lijken nauw gekoppeld te zijn aan het spraakproces. Ten eerste maken mensen bijna uitsluitend gebaren tijdens het spreken (en bijvoorbeeld niet tijdens het luisteren) en ten tweede lijkt de betekenis van die gebaren, voorzover deze te achterhalen valt, direct gerelateerd te zijn aan de spraak. Het onderwerp van deze dissertatie is de productie van gebaar en spraak, en hun onderlinge relatie.

De gebaren van een spreker zijn anders dan de gebaren die doven gebruiken om met elkaar te communiceren. De gebarentaal van doven is een echte taal. Deze gebarentaal heeft, net als bijvoorbeeld Chinees of Nederlands, syntactische, semantische, en morfologische regels die bekend zijn bij de sprekers van die taal. Voor de meeste gebaren die sprekers maken bestaan dit soort regels niet.

Gebaren kunnen op veel verschillende manieren worden gecategoriseerd. De hier gevolgde typologie is van McNeill (1992). Deze typologie onderscheidt de volgende categorieën. *Deictische* gebaren zijn (wijs)gebaren die een locatie of een bepaalde richting aangeven. In de Nederlandse en Engelse taal wordt voor deictische gebaren meestal de wijsvinger gebruikt, maar ook andere vingers, het hoofd of de lippen worden in sommige culturen gebruikt om deictische gebaren te maken. Daarnaast zijn er gebaren die aangeduid worden met het Engelse woord *emblems*. Dit zijn gebaren met een voor de gebruikers van een taal bekende betekenis. De duim omhoog om aan te geven dat iets "OK" is, of de wijsvinger naar de lippen brengen om iemand tot stilte te manen zijn voorbeelden van *emblems*. Verder kunnen we *beats* onderscheiden. Dit zijn ritmische op en neer bewegingen van de hand die geen betekenis lijken te hebben. Het vermoeden bestaat dat het ritme van deze gebaren gerelateerd is aan de fonologische kenmerken van de tegelij-

kertijd geproduceerde spraak, maar daarvoor is nog geen overtuigend bewijs geleverd. Tot slot is er de belangrijke categorie van de *iconische* gebaren. De vorm van deze gebaren heeft een betekenisrelatie met het onderwerp van de spraak. Met de wijsvinger een spiraalvormige beweging maken als men het over een draaikolk heeft is een voorbeeld van een iconisch gebaar.

De relatie tussen spraak en gebaar wordt vaak onderzocht door videoopnamen van sprekers te maken en deze aan een nauwkeurig onderzoek te onderwerpen. Deze methode heeft veel kennis opgeleverd over de vele manieren waarop gebaar en spraak gerelateerd zijn. Het nadeel van deze methode is echter dat het vrijwel ondoenlijk is om algemene uitspraken te doen over gebaar en spraak. De oorzaak hiervan ligt in de grote variabiliteit van gebaren. Verschillende sprekers maken verschillende gebaren, maar zelfs één en dezelfde spreker maakt vaak verschillende gebaren, ook al is het onderwerp van de spraak hetzelfde. Een ander probleem is dat veel gebaren spontaan en onbewust gemaakt worden. Het heeft dus geen zin om mensen dit soort gebaren "op commando" te laten maken. Dit methodologische probleem treedt voornamelijk op bij het onderzoeken van iconische gebaren en beats. Bij *emblems* en deictische gebaren kan men, dankzij het feit dat deze gebaren een voorspelbare vorm hebben, gecontroleerde experimenten uitvoeren.

Psychologisch onderzoek naar gebaar en spraak richt zich voornamelijk op de vragen *waarom*, *hoe*, en *wanneer* gebaren gemaakt worden.

In hoofdstuk 2 wordt de vraag *hoe* mensen gebaren maken benaderd door een blauwdruk voor een informatieverwerkingsmodel te formuleren. Een van de voordelen van het formuleren van een model is dat men een overzicht krijgt van de verwerkingsoperaties (berekeningen) die noodzakelijk zijn voor het produceren van gebaar en spraak. Het model dat wordt geformuleerd, het "Sketch Model", is een uitbreiding en aanpassing van Levelt's (1989) model voor het spreken. In het "Sketch Model" wordt ervan uitgegaan dat het maken van gebaren, net als spraak, een communicatieve functie heeft. Tijdens het voorbereiden van de spraak wordt ook reeds het begeleidende gebaar voorbereid. Daarna worden volgens het model spraak en gebaar grotendeels onaf-

hankelijk van elkaar verwerkt. Dit is een ander uitgangspunt dan dat van Krauss et al. (1996), die ervan uitgaan dat gebaren een *epifenomeen* van het spreken zijn. Volgens Krauss et al. worden gebaren niet gemaakt om informatie toe te voegen aan de spraak, maar om het spraakproces zelf te ondersteunen. Dit van het Sketch Model verschillende uitgangspunt over de functie van gebaren leidt dan ook tot een geheel ander model. McNeills (1992) *growth point* theorie gaat, net als het Sketch Model, wèl uit van het idee dat spraak en gebaar samenwerken om te communiceren. Omdat echter de growth point theorie niet in termen van informatieverwerking is geformuleerd, is het erg moeilijk om toetsbare voorspellingen te doen. Het Sketch Model doet wel een aantal duidelijke voorspellingen, met name over de vraag *wanneer* mensen gebaren, in relatie tot de geproduceerde spraak.

In hoofdstuk 3 wordt een theorie van Kendon (1980) getoetst, die luidt dat gebaren gelijktijdig met of voorafgaand aan de "intonational peak syllable" van de gerelateerde spraak gemaakt worden. De "intonational peak syllable" is de lettergreep uit een intonatie-eenheid de het meest geaccentueerd is. In de verrichte experimenten werden de proefpersonen verzocht plaatjes te beschrijven en ernaar te wijzen. Door het registreren van de spraak en de bewegingen van de wijzende hand kon de temporele relatie tussen gebaar en spraak nauwkeurig worden onderzocht. In het eerste experiment bleek dat de gemaakte gebaren absoluut niet beïnvloed werden door wanneer de beklemtoonde lettergreep van een woord werd uitgesproken. Of proefpersonen nu "de krokoDIL" zeiden, of "de CAmera", het had geen invloed op de *timing* van het gebaar. Het toevoegen van intonatie aan de spraak, door het laten benadrukken van ofwel de kleur ofwel de naam van het getoonde object, leidde wel tot een adaptatie van het gebaar: naarmate de geaccentueerde lettergreep later werd uitgesproken werden de wijsgebaren langzamer en langer. Het analyseren van een aantal spreekfouten die door de proefpersonen werden gemaakt leverde een nog duidelijker adaptatie-effect op. Als de spraak haperde, paste het gebaar zich vrijwel onmiddellijk aan, zodat de uiteindelijke temporele relatie tussen gebaar en spraak weer hetzelfde was als bij spreken zonder hapering.

Hoewel een aantal bevindingen uit deze studie het Sketch Model on-

dersteunen, lieten de resultaten zien dat de interactie tussen gebaar en spraak veel hechter is dan in dat model wordt verondersteld. In hoofdstuk 5 wordt het model daarom aangepast. In het aangepaste Sketch Model wordt verondersteld dat de spraakprocessen pas kunnen beginnen als het deelproces dat verantwoordelijk is voor het plannen van het gebaar (de *gesture planner* in het model; zie figuur 5.1) klaar is, en een signaal stuurt aan het deelproces dat het spraakproces initieert (de *conceptualizer*). Deze aanpassing aan het model verklaart tevens op een eenvoudige manier het fenomeen dat gebaren meestal iets voorafgaan aan de gerelateerde spraak.

In hoofdstuk 4 wordt tenslotte de vraag *waarom* mensen gebaren maken onderzocht. Een aantal auteurs stelt dat iconische gebaren niet communicatief bedoeld zijn, maar slechts dienen om het spreken te vergemakkelijken. Dit zou de reden kunnen zijn dat mensen tijdens telefoongesprekken ook gebaren maken, ook al kan de luisteraar hun gebaren niet zien. De experimentele resultaten die met name Krauss en zijn collega's aanvoeren als bewijs voor hun theorie kunnen echter allemaal worden verklaard met de eenvoudige assumptie dat gebaren een communicatieve functie hebben. De vraag of gebaren het spraakproces ondersteunen staat bovendien los van de vraag of zij wel of niet communicatief bedoeld zijn. In de experimenten in hoofdstuk 4 worden twee hypothesen over de mogelijke spraakondersteunende functie van gebaren getoetst. De eerst hypothese is dat het maken van gebaren de geheugenrepresentaties die aan het gebaar ten grondslag liggen gemakkelijker toegankelijk maakt. Dit wordt de *retrieval* hypothese genoemd. De tweede hypothese is dat het maken van gebaren het proces van het spreken zelf faciliteert, bijvoorbeeld doordat het maken van gebaren het gemakkelijker maakt de juiste woorden of concepten op te halen uit het geheugen. Deze laatste hypothese wordt de *encoding* hypothese genoemd.

Om deze twee hypothesen te toetsen werd de proefpersonen verzocht plaatjes met een aantal geometrische figuren te beschrijven zodat een andere proefpersoon deze kon tekenen. De beschrijver en de tekenaar konden elkaar in deze experimenten niet zien, om te verhinderen dat gebaren om communicatieve redenen gemaakt werden. Bij de helft van de

plaatjes diende de proefpersoon de plaatjes eerst uit het hoofd te leren alvorens aan de beschrijving te beginnen. Bij de andere helft werden de plaatjes direct van het scherm beschreven. De plaatjes zelf waren verdeeld in enerzijds een groep die gemakkelijk te beschrijven was, en anderzijds een groep die heel moeilijk te beschrijven was. De laatste groep plaatjes was moeilijk te beschrijven omdat het erg lastig was om te vertellen waar de deelfiguren van de plaatjes ten opzichte van elkaar geplaatst waren. Het resultaat van dit experiment was dat proefpersonen meer gebaren maakten als zij de plaatjes uit het hoofd moesten leren, dan als ze de plaatjes konden zien. Dit resultaat ondersteunt de *retrieval* hypothese. Er was echter geen verschil in de frequentie van de gebaren tussen de moeilijke en de gemakkelijke plaatjes, zelfs niet in een vervolgexperiment dat speciaal gericht was op het detecteren van een eventueel effect van de moeilijkheid van de plaatjes. Hoewel het gevaarlijk is om uit een nuleffect vergaande conclusies te trekken, lijkt deze bevinding toch tegen de *encoding* hypothese in te gaan. Als de *encoding* hypothese klopt, dan is het onwaarschijnlijk dat het grote verschil in moeilijkheid tussen de twee groepen plaatjes geen enkel effect zou hebben op de gebaarfrequentie.

De bevindingen van de experimenten uit hoofdstuk 4 zijn in overeenstemming met het Sketch Model. In dat model wordt immers verondersteld dat interacties tussen gebaar en spraak alleen in een vroeg stadium (in de *conceptualizer*) mogelijk zijn. De *retrieval* hypothese is daardoor in overeenstemming met het Sketch Model. De *encoding* hypothese is dat niet, omdat deze veronderstelt dat het maken van gebaren het spreken helpt op het niveau van de *formulator*, hetgeen niet mogelijk is in het Sketch Model.

Een korte samenvatting van de belangrijkste bevindingen in deze dissertatie kan als volgt gegeven worden. Ten eerste is gevonden dat spraak en gebaar temporeel nauw gekoppeld zijn. Het tijdsverloop van de spraak heeft invloed op het tijdsverloop van het gebaar, en andersom. Verder speelt ook de intonatie van de spraak een rol in het tijdsverloop van het gebaar: Kendons (1980) claim dat gebaren gesynchroniseerd zijn met de "intonational peak syllable" in de spraak is bevestigd. Een andere bevinding is dat het maken van gebaren het spreken kan ondersteunen

door het (her)activeren van visuele geheugenrepresentaties in het korte termijn geheugen. Tot slot kan gesteld worden dat de productie van gebaar en spraak, en hun interactie, dermate complex zijn dat het gebruik van een informatieverwerkingsmodel onontbeerlijk is voor het doen van toetsbare voorspellingen.

CURRICULUM VITAE

Jan-Peter de Ruiter was born in Leiden in 1964, a city to which he returned in 1982 to get a taste of computer science and a variety of social sciences at Leiden University. In 1987 he moved to the city of Nijmegen to work as a computer programmer for the laboratory of the University Hospital. From 1988 to 1992 he studied cognitive science at the Catholic University of Nijmegen. After finishing his studies at the K.U.N., he worked again as a computer programmer, this time at Leiden University. In 1994 he received a stipend from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften to do a dissertation project on gesture and speech at the MPI for Psycholinguistics. Currently, he is continuing his gesture research at the MPI as a post-doc.

EMPTY PAGE

EMPTY PAGE