When Hands Talk to Mouth. Gesture and Speech as Autonomous Communicating Processes

Hannes Rieser
Collaborative Research Center
“Alignment in Communication” (CRC 673)
Bielefeld University, Germany
hannes.rieser@uni-bielefeld.de

Abstract

The implementation of speech-gesture interfaces is one of the vital problems in formal research on multi-modal discourse. This paper provides empirical evidence that, due to asynchronous occurrences of gesture and speech, speech-gesture interfaces cannot be expressed in purely static structural terms resting on a speech-gesture map. As an alternative, a methodology is suggested which models gesture and speech as independent communicating processes generating together multi-modal content. It is based on a dynamic process algebra, the π-calculus. To meet the descriptive needs of speech-gesture interface construction, the π-calculus is extended to a hybrid λ-π-calculus devised to handle higher order information.

1 Introduction

Recently, there has been a growing interest to investigate the coordination of gesture and speech in multi-modal discourse, originally initiated by scholars like McNeill (1992) and Kendon (2004). I’ll take up this topic in my paper as well, providing a new approach. The leading idea, motivated by corpus studies in sec. 5, is that speech and gesture work as independent processes, abstract agents which communicate and together produce a multi-modal content, e.g., if a winding gesture modifies the word “street”. It will be shown in due course that this also necessitates a move from classical algorithmic modelling, be it λ-calculus, Montague Grammar (MG) or some brand of Dynamic Semantics, to process modelling using dynamic calculi such as the π-calculus. I will substantiate this idea in the following way: Sec. 2 starts with some assumptions embodying the idea to take gesture and speech as independent processes. This is further motivated in sec. 3. Sec. 4 presents McNeill’s influential observations on speech-gesture coordination. In sec. 5 examples of static speech-gesture interfaces are provided resting mainly on McNeill’s ideas. Sec. 6 elaborates on three case studies showing asynchrony of gesture and speech yielding counter examples to static speech gesture interfaces. Sec. 7 comes with intuitive process analyses for the asynchrony cases involving parallel processes and process-interaction. In sec. 8, I introduce a process algebra, namely π-calculus, show how to extend it to a hybrid of λ-π-calculus, and use the resulting machinery for the description of gesture-speech interaction. I close with some indications for future research in sec. 9.

2 Assumptions

I assume that speech and gesture have meaning, say along the lines of a Peircean semiotics. As a consequence, I take it that speech meaning and gesture meaning can be represented and computed independently but that there is some coordination between them. This is obvious, e.g., from demonstrations accompanying the use of indexical expressions: demonstrations have to be coordinated with the production of the indexical (Lücking et al., 2015). The *locus* of speech-gesture coordination is informally called “interface” here, following common practice in software engineering to compute information of different type from different sources. In the interface, speech information and gesture information are stored and processed.

3 Idea of the paper

As is obvious from the remarks on the interface, the interface between speech meaning and gesture meaning has to be expressed in a formal way. The question is then which formalism to use. The options are data structures suitable to interface information. So it is no surprise that Mc-

Neill (1992) used frame-structures (reconstructed in Röpke (2011)). A more recent concept resorts to AVMs in an HPSG-representation (Lücking, 2013). In general, I assume that the modelling of gesture must be based on rigid (rated) annotation, annotation playing for gestures the same role as syntax representation plays for linguistic utterances. Speech acts and gesticulations are widely different types of structure, there is no natural mapping from one to the other comparable to a syntax-semantics-map. As will be shown in the case studies below (sec. 5), natural speech gesture interface data resist modelling in conventional structural terms (such as trees, AVMs or pure FOL-representations). As a consequence, so I argue, one must look for different conceptualizations. The main problems for a natural mapping from gesture to speech are that the gesture often does not exactly overlap its fitting speech counterpart: It comes too early, too late or extends over too much language material. So, there is no semantic synchrony. A machinery which seems to be able to capture this dynamics at least partially are process algebras such as the π-calculus (Parrow (2001) and Sangiorgi and Walker (2001)), the Calculus of Communicating Systems (CCS, Milner (1999)) or Communicating Sequential Processes (CSP, Hoare (1985)). Using one of these will move one from an object and proposition metaphysics to one based on processes.

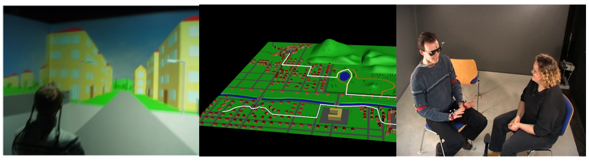


Figure 1: The virtual town, the route traversed, Route-giver and Follower sitting in the cave.

The dynamics of the speech-gesture relation will be shown using the speech and gesture alignment corpus, SaGA (Lücking et al., 2013). It contains 25 route description dialogues from three camera perspectives. The dialogue participants are a Route-giver and a Follower, the Route-giver explaining his/her route through a virtual town to a Follower. Lücking et al. gathered video and audio data, body movement tracking data, and eye-tracking data. Approx. 7500 gestures have been identified, 6000 of them annotated and rated. Due to the experimental setting, they have to deal with the genre of multi-modal task-oriented dialogue with many specific dialogue structures, such as clarification sequences, repetitions and tests.

4 Mc Neill on Speech-gesture Coordination

McNeill (1992) using the so-called Tweety-data was the first scholar to provide generalizations on speech-gesture coordination which are widely used for interface construction, although, as I will show below, his approach is in the end too normative and prone to falsification. Here I provide McNeill's semantic synchrony rule (McNeill (1992), p. 27) and his definition of stroke (McNeill (1992), p. 83) for further use.

Semantic synchrony rule: Semantic synchrony [of gesture and speech, author] means that the two channels, speech and gesture, present the same meanings at the same time. The rule can be stated as follows: "if gesture and speech co-occur they must cover the same 'idea unit'" [i.e., content, author].

Stroke: Stroke [. . .] is the peak effort in the gesture. It is in this phase that the meaning of the gesture is expressed. The stroke is synchronized with the linguistic segments that are co-expressive with it.

From McNeill's definitions of semantic synchrony and stroke it follows that the set-up of a speech-gesture interface is provided by (the content of) the gesture's stroke and the meaning of the synchronised linguistic material. These two have to interact. For example, an iconic gesture indicating a square can interact with the semantics of, say, "envelope", indicating the envelope's shape.

5 Static speech-gesture interfaces: frames and HPSG-matrices

McNeill (1992) was interested in specifying the generation of speech-gesture ensembles as shown in fig. 2. The important issue here is that a filled frame is used to store the information necessary for generating speech and (optionally) an accompanying gesture. We get the information needed packed into one static data structure. A more recent variant of a static data structure is provided by Lücking (2013) who uses an HPSG-grid to model speech-gesture interfaces (fig. 3).

The relation under discussion on this grid is a two-dimensional "round", "round2", which gets

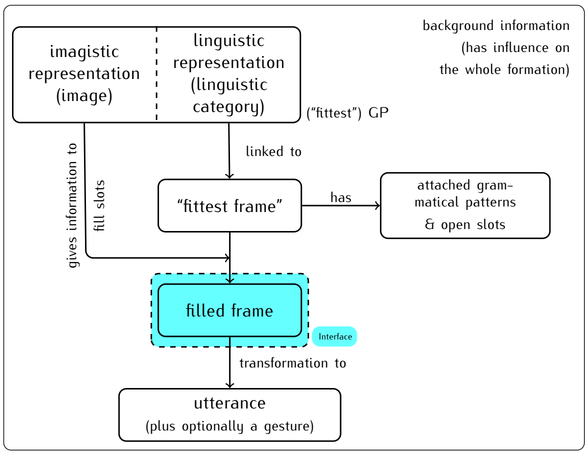

imagistic representation (image)
linguistic representation (linguistic category)
background information (has influence on the whole formation)
("fittest") GP
linked to
gives information to
fill slots
"fittest frame"
has
attached grammatical patterns & open slots
filled frame
Interface
transformation to
utterance (plus optionally a gesture)

Figure 2: McNeill's speech-gesture generation frame as reconstructed in Röpke (2011).

$$\begin{bmatrix} \textit{sg-ensemble} & \\ \text{CONT} & \boxed{3}\begin{bmatrix}\text{INDEX} & i \\ \text{RESTR} & \left\langle \boxed{5}\begin{bmatrix}\textit{form-pred} & \\ \text{RELN} & \text{round2} \\ \text{ARG} & i \\ \text{CVM} & \begin{bmatrix}\text{VEC} & \text{axis-path}(i,\mathbf{v}) \\ \text{PATH} & \boxed{7}\,(360^\circ(\mathbf{v}))\end{bmatrix}\end{bmatrix}\right\rangle\end{bmatrix} \\ \text{S-DTR} & \begin{bmatrix}\textit{verbal-sign} & \\ \text{PHON} & \boxed{1}\begin{bmatrix}\text{ACCENT} & \boxed{6}\end{bmatrix} \\ \text{CONT} & \boxed{3}\end{bmatrix} \\ \text{G-DTR} & \begin{bmatrix}\textit{gesture-vec} & \\ \text{AFF} & \left\langle\begin{bmatrix}\text{PHON} & \boxed{1}\begin{bmatrix}\text{ACCENT} & \boxed{6}\,\textit{marked}\end{bmatrix}\end{bmatrix}\right\rangle \\ \text{TRAJ} & \boxed{7}\,\mathbf{UP} \oplus -\mathbf{RT} \oplus -\mathbf{UP} \oplus \mathbf{RT} \\ \text{CONT} & \begin{bmatrix}\text{MODE} & \text{exemplification} \\ \text{EX-PRED} & \boxed{5}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

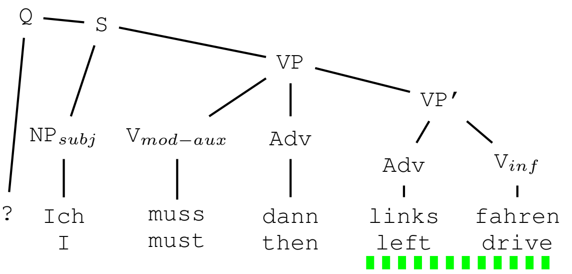
Figure 3: HPSG-grid to model speech-gesture interfaces from Lücking (2013), p. 249.

its semantics from a TRAJectory in the G(esture)-DaughTeR as is evident from the unification $\boxed{7}$. As in the frame case, we have a static structure. The trajectory's semantics can enter into exactly one position of the RESTRiction of "round2". The set-up of the speech-gesture ensemble is quite powerful due to unification but we cannot go into details here, check especially $\boxed{3}$ and $\boxed{5}$. The content of fig. 2 and fig. 3, respectively, might well serve as a kind of *explicans* to the McNeill quotes above. Mainly for didactic reasons I have chosen Lücking's approach as a prototypical one here but I think that the same arguments apply tot he HPSG-based speech gesture interfaces of Alahverdzhieva and Lascarides (2010) who also use structure-based technologies. Furthermore, work in the SDRT-tradition focusing on the explication of gesture meaning is based on similar interface conceptions (Lascarides and Stone (2006) and (2009)) and faces similar falsifying instances. Hopefully, the difference to the process-based proposals made in this paper will become clear from the following case studies (sec. 6) and the process analyses in sec. 7. How the findings presented here carry over to incremental theories of information in the manner of, e.g. Hough et al. (2015), still remains to be investigated, however, the suspicion is that they do carry over. Turning to the point of view of hypothesis falsification the question arises whether we find gesture-speech occurrences where speech and gesture belong together intuitively but do not obey McNeill's synchrony rule. Below I present the essentials of three case studies showing exactly such falsifying instances. They also serve as falsifying instances for static speech-gesture interfaces.

6 Three case studies: Asynchrony of gesture and speech (based on Hahn and Rieser (2012))

In the following, intuitive notions like "channel", "communicate", "interaction", "interfacing", "process" or "sending" are used. They are given a proper algorithmic reconstruction in sec. 8.

6.1 Case I: Indexing is held too long



(a) Syntax of Follower's clarification request. The stroke is marked with a green dashed line.

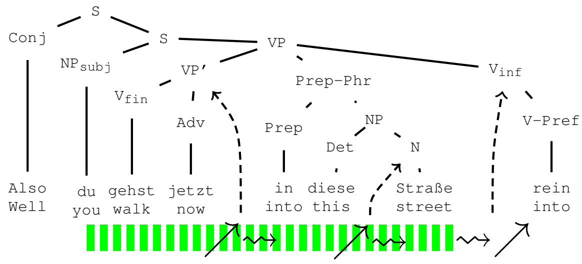
$$\begin{bmatrix}\text{Hand\_Shape} & L \\ \text{Back\_Of\_Hand\_Direction} & BTL \\ \text{Palm\_Direction} & PTB \\ [\ldots] & \end{bmatrix}$$

(b) AVM of gesture annotation.

Figure 4: Datum of Case I.

Fig. 4 shows a Route-giver and Follower exchange, the syntax of the clarification request *?I must then left drive* and the annotation of the Follower's gesture, a demonstration to the left. The green marks indicate the gesture stroke overlap with *left* and *drive*. According to McNeill the stroke should only overlap with *left*. Hence, the Follower's indexing is held too long. At first sight an explanation could be given which is in agreement with McNeill, namely, if we interface the gesture stroke with the VP'. However, doing that we would lose bottom-up compositionality, because the terminal "left" is not related to the stroke.

6.2 Case II: Object gesture must wait to compose



(a) Syntax of Follower's clarification request. The stroke is marked with a green dashed line.

Hand_Shape	*B-spread*
Back_Of_Hand_Direction	*BAB/BTL>BTL>BAB* >BAB/BTL
Palm_Direction	*PTB/PTL>PTB>PTL* >PTB/PTL
Wrist_Movement_Direction	*MF/ML/MU*

(b) AVM of gesture annotation.

Figure 5: Datum of Case II. Stroke of gesture overlapping several constructions, *inter alia*, the subject NP.

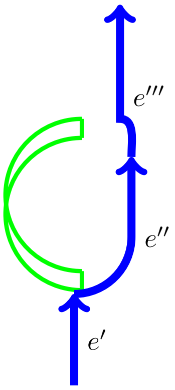
Fig. 5 shows a clarification request of a Follower. Again I provide the gesture annotation and the syntax structure, here of the 2^{nd} utterance. The green marks represent the stroke of the winding gesture. Observe that the winding information is not contained in the utterance, so we have additional information in the gesture. Although there are several options for speech-gesture interfaces, the preferred *locus* of integration is "street", yielding winding street in a multi-modal way, whereas it could not easily be combined with "walk now" or "into". Since the stroke starts overlapping with "you", we have again a counter-example to McNeill's synchrony rule.

6.3 Case III: Multi-parallelism and anaphora

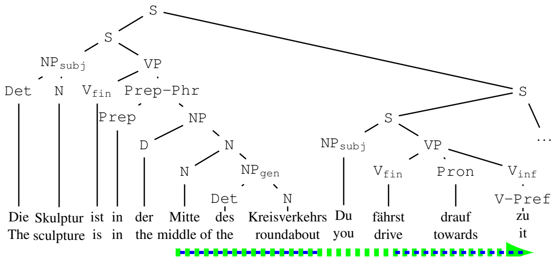
```
Es ist aber    auch ein Kreisverkehr.
It is  however also a   roundabout.
Die Skulptur  ist in der Mitte  des
The sculpture is  in the middle of the
Kreisverkehrs. Du  fährst drauf   zu,
roundabout.    You drive  towards it,
rechts herum     und dann  ja    und
right  around it and then, well, and
dann quasi geradeaus      sozusagen
then quasi straight-ahead so to speak
die  abbiegen.
this one leaving.
```

(a) Route-giver's directive.

LH	
Hand_Shape	*C-loose*
Back_O_H_Direction	*BAB*
Palm_Direction	*PTR*
RH	
Hand_Shape	*H-loose*
Back_O_H_Direction	*BAB*
Palm_Direction	*PDN*
Wrist_Mov_Direction	*MF*



(b) LH- and RH-annotation and Trajectories e', e'', e''' representing the trajectories of "drive towards", "right around it", and "leaving it", respectively.


S
S
NP$_{subj}$
VP
Det
N
V$_{fin}$
Prep-Phr
S
Prep
NP
D
N
S
NP$_{subj}$
VP
N
NP$_{gen}$
V$_{fin}$
Pron
V$_{inf}$
Det
N
V-Pref
...
Die Skulptur ist in der Mitte des Kreisverkehrs Du fährst drauf zu
The sculpture is in the middle of the roundabout you drive towards it

(c) Syntax tree and stroke overlaps as dashed lines, left hand green, right hand blue.

Figure 6: Datum of Case III.

Fig. 6 represents the Route-giver's utterance, the annotation of the left hand, the right hand and the trajectories e', e", and e"'. The "natural interface" in these cases is not marked by a gesture-stroke speech overlap. A more elaborate description of the right-hand and the left-hand activities and their relation to speech will be given in the next section.

6.4 Generalisation

Let us generalise from the findings in the case studies: Given that we have at least two information channels, an alternative to static speech-gesture interfaces emerges: We model the respective information on two channels and how they communicate. Still, after having dealt with the issues due to the falsification, we must be aware of the fact that at the ultimate speech gesture contact points, i.e., if we have successfully singled out the appropriate interface, we will encounter problems as those indicated by McNeill, Lücking, Rieser (2013) and others, namely, how to represent the alignment of speech meaning and gesture meaning. This shows that these researchers discovered something important but used an idealised case prototypes of which can also be found in the data.

7 Process analyses for the asynchrony cases

This section contains intuitive analyses of the case studies. They are informally expressed and serve as a sort of precondition for the discussion about communicating processes in sec. 8.

7.1 Taking up case I: Indexing is held too long

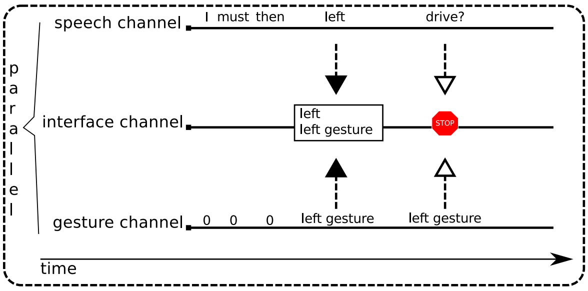


Figure 7: Three parallel channels: RH, LH, interface and speech channel.

Fig. 7 depicts three channels operating in a parallel way. On the speech channel we have the utterance “I must then left drive”?. On the gesture channel there are first empty events indicated by 0. The interface channel encodes the interaction between the information on the speech channel and the information on the gesture channel. It also indicates where the interfacing can occur (boxed area) and where it can’t. Accordingly, the semantics of the word “left” and the left gesture interact in some time interval but there is no interaction afterwards, indicated by the red stop sign. So, the idea is to restrict the activity of the gesture, more precisely, that of its semantic representation, to the word “left”. Observe that the gesture meaning itself is in no ways annihilated; it remains on the gesture channel.

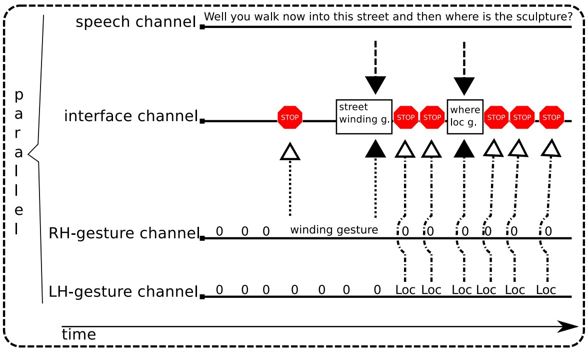


Figure 8: Four parallel channels: RH, LH, interface and speech channel.

7.2 Taking up case II: Object gesture must wait to compose

As fig. 8 shows, we have a RH- and a LH-gesture channel, both interacting with the speech channel which transports “Well you walk now into this street and then, where is the sculpture”?. Here the RH’s gesture comes too early at “now” which cannot combine with the winding. Since it continues to send *via* the extended gesture stroke, it can finally cooperate with “street” yielding in the end the multi-modal semantics ⟦winding street⟧. The LH gesture starts to communicate when speech contributes “and” and “then” on the speech channel. However, to become effective, it has to wait until “where” turns up, then providing indexical information for it in the gesture space (producing Quinean deferred reference). After that the speech’s cooperation potential is used up and the LH-gesture is barred off from contribution to the interface channels.

7.3 Taking up case III: Multi-parallelism and anaphora

Fig. 9 shows that we have various interfaces active. LH and RH first communicate to produce a cylindrical shape and the shape information then tries to get access to the speech level. The example also shows the use of a linguistic anaphora “the sculpture” and of an anaphora at the gesture level. More on that further down. In more detail,

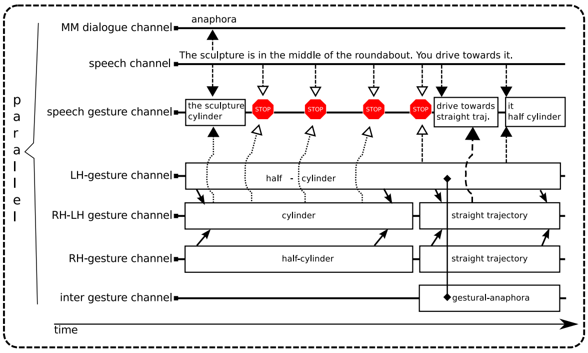


Figure 9: RH- and LH-gesture channel communicate forming the "composite" RH-LH gesture channel. This channel communicates with the speech channel on the speech-gesture channel. In addition, there is a multi-modal dialogue channel on which linguistic anaphora ride and an inter-gestural channel for gestural anaphora.

the LH-gesture and the RH-gesture each form a half-cylinder and together shape a cylinder. The cylinder-information can communicate with the speech information "the sculpture" on the speech-gesture channel. Then the cylinder information is stopped from interacting with speech. Afterwards the LH and the RH part company. The RH shapes a straight trajectory whereas the LH still signs the half-cylinder, forming a "gestural anaphora" for the whole cylinder. However, LH and RH start to cooperate on the RH-LH-interface channel. Together they indicate a situation involving a cylindrical object (LH) and a path around it (RH) which can contribute to the meaning of "drive towards" which clearly involves a target. The LH continues to send information, contributing after at least one stop indexical meaning to the anaphora "it".

7.4 Evaluation of data and development of the formal mechanisms needed to describe flexible speech gesture-interaction

If we look at the speech gesture interaction, we find that actions like "stop, I do not want information" (indicated by the red stop sign), processes and process interactions seem to be the most basic entities. We encountered different types of processes, empty ones, speech-gesture, gesture-gesture. Processes run in parallel as our timelines indicate. They hook up to each other *via* interfacing. They emit or receive information. In the case studies gesture is as a rule the emitting source and speech the receiver. Receiving can imply that processes are changed by information, remember the multi-modally specified winding street. However, information can also be neglected or blocked. Processes can be recursive, this can be seen, when a process tries to communicate several times (thus generating a daughter process of the same type) but is barred from the interface. Interactions among channels come in sequences. Clearly, we need an algorithm which can capture at least some of that.

8 From λ- to π-calculus. The step to process algebra

Before we deal with processes, we enter a familiar field: the λ-calculus. Formal work in NL semantics often relies on applied λ-calculus. It has logical constants, constants for individuals and relations, operators for all styles of variables plus the λ-operator. It often works with a generalised quantifier representation and rules of $\alpha\beta\eta$-conversion (see Curry et al. (1974), p. 92). It has inspired semantic work from Church and Curry to Montague and beyond. In contrast, the π-calculus' basic entities are names, represented by lower case letters. They are used by processes/channels ("channel" now being a technical term) for interactions. Interactions have to be formally indicated: Capabilities for actions are provided by so-called *prefixes* π:

$$\pi \ := \ \overline{x}y \mid x(z) \mid \tau \mid [x = y]\pi$$

Then we have processes P and summations M:

$$P \ := \ M \mid P \mid P' \mid \nu z P \mid !P$$

$$M \ := \ 0 \mid \pi.P \mid M \ + \ M'$$

Among the prefixes we have an output prefix $\overline{x}y$ and an input prefix $x(z)$. $\tau.P$ evolves invisibly to P. There is a match prefix, $[x = y]$, in $[x = y]\pi.P$. In a sum $P + P'$ either P or P' can be executed but not both. "$P|P'$" is called "composition". In such a composition, P and P' can be executed independently, in parallel or interact *via* shared names, yielding output-input devices. Shared names are already indicated in π above. $\nu z P$ states that the scope of name z is restricted to process P, in traditional parlance, z is treated much like a bound variable. $!P$ denotes infinite iteration, defined as $P|!P$, i.e., a process composed with an iteration of processes.

Given our intuitive analyses of asynchrony cases in sections 6 and 7, what do we get from the π-calculus? First of all, a technical nomenclature for our intuitive distinctions like process etc. (see the list in sec. 7) and then algorithmic means for modelling them. In more detail: We have parallel channel modelling *via* composition "|". As already indicated, there are output-input devices *via* the prefixes $\overline{x}y$ (outputting y) and $x(z)$ (receiving a name *via* x and substituting it for z in the subsequent process). We have types of binding, tests and arbitrarily deep recursion due to replication. In addition, type systems for channel names can be given, a device which we will exploited below.

8.1 Typing and a hybrid λ-π-machinery

So, the advantages of π seem to be fairly clear. But hold on! Essentially, we would like to have the expressive power of the higher order λ-calculus in the interfaces, gesture-gesture and gesture-speech alike, as we have seen them in the asynchrony studies. The reason is that the information seems to be higher order. My suggestion is to achieve that through

(a) using λ-operator and π-operator based definitions for λ-calculus names a, b, etc. resulting in mixed λ-π-expressions

(b) using typed λ-calculus constants as π-calculus names for channels, λ-calculus constants are given the status of π-calculus names

(c) letting channels have a MG type such as e, $\langle e, t\rangle$, $\langle\langle e, t\rangle, \langle e, t\rangle\rangle$, etc. in order to match them with type-fitting names.

I call the π-calculus extended with (a), (b), (c), hybrid λ-π-calculus.

We are now equipped to handle interfacing speech-gesture processes and turn to an illustrative example. We do the relatively simple case I from section 7, indexing is held too long (cf. fig. 4). For didactic reasons, I use the English word by word translation "I must then left drive?" and reconstruct the crucial "left drive" in λ-π-calculus terms. A further simplification is added: Because of operators "must" and "then" I use a more tractable version of the English translation, namely "Must then I left drive?" I provide first the speech representation, then the speech interface representation and the definitions of names and types for the λ-π-calculus representation (cf. table 1). Holding indexing is modelled by "$!P$". Parallel channels of speech and gesture are modelled by "|".

Speech representation	Speech interface representation	Types and definitions for π-calculus names
$I := \lambda P.P(I)$	same	$\overline{x'}, x', \overline{z}, u, w$: MG-type $< e, t >$
drive $:= \lambda x.\text{drive}'(x)$	same	$\overline{x}, x, y$: MG-type $<< e, t >, < e, t >>$
left $:= \lambda f \lambda z.\text{left}'(f)(z)$	$\lambda z.\text{left}'(w)(z) \ \wedge\ y(w)(z)$	$a := \lambda z.\text{left}'(w)(z) \ \wedge\ y(w)(z)$
must $:= \lambda p.\Box p$	same	
then $:= \lambda p \lambda q.\text{then}'(p, q)$	same	
Gesture representation	Gesture interface representation	
$!\overline{x}\text{left}'.0$	same	
Speech-gesture interface	gesture representation	speech representation
$x(y).x'(w).\ a.0 \mid$	$!\overline{x}\text{left}'.0 \mid$	$\overline{x'}\text{drive}'.0$

Table 1: Speech representation, gesture representation and their λ-π-interface.

As in the figures, green colouring indicates gesture information. Replication definition for $!\overline{x}\text{left}'.0$ yields:

$$x(y).x'(w).\ a.0 \mid \overline{x}\text{left}'.0 \mid !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \qquad (1)$$

We substitute π-names in the processes by their λ-π-definitions and get:

$$x(y).x'(w).\lambda z.\text{left}'(w)(z) \ \wedge\ y(w)(z).0 \mid \overline{x}\text{left}'.0 \mid !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \qquad (2)$$

$\overline{x}$ outputs left$'$ to x: y is instantiated with left$'$:

$$x'(w).\lambda z.\text{left}'(w)(z) \ \wedge\ \text{left}'(w)(z) \mid 0 \mid !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \qquad (3)$$

$\overline{x'}$ outputs drive$'$ to x': w is instantiated with drive$'$:

$$\lambda z.\text{left}'(\text{drive}')(z) \ \wedge\ \text{left}'(\text{drive}')(z) \mid 0 \mid !\overline{x}\text{left}'.0 \mid 0 \qquad (4)$$

We get the property "$\lambda z.\text{left}'(\text{drive}')(z) \ \wedge\ \text{left}'(\text{drive}')(z)$" which after normalization becomes "$\lambda z.\text{left}'(\text{drive}')(z)$".

Observe that "compositionally used up" information results in 0-events. For some reasons (perhaps to facilitate coherence, to add emphasis or to maintain the focus) the gesture is kept on its channel. This is what "$\overline{x}\mathrm{left}'.0$" expresses. Again, this is additional information for separating the gesture channel and the speech one. So what we get in the end is an algorithmic rendering of the intuitive representation in fig. 7.

So, the gesture does not contribute new content to the speech content. But, while the word "left" evaporates, the indexing on the gesture channel is still visible as we have it in the datum and the diagram fig. 10. It can still be SEEN when the next word "drive" is already HEARD, leading to a division of labour among channels.

8.2 A note on generalisability

Finally, a word on generalisability of the λ-π-calculus account might be in order: We need multi-channel renderings in various multi-modal contexts anyway, take, e.g., tone-group information not strictly co-extensive with syntax trees, the information contained in eye-tracking data or in body postures. So, multi-channel representations seem to be an imperative research venue to follow.

9 Conclusion and further research

As shown in the case studies, in the SaGA data speech portions and gesture strokes do not perfectly synchronize. We have seen that grammar-based speech gesture interfaces cannot deal with gestures produced too early, lagging behind or intruding" into "alien" speech material by, e.g., crossing propositional boundaries, expressing contradictory content etc. As a way out we propose to consider speech and gesture as autonomous concurrent processes communicating with each other *via* an interface. This can be achieved by exploiting the facilities of the suggested λ-π-calculus to model higher order properties of concurrent speech-gesture events and gesture-gesture events.

As the λ-π-hybrid shows, we have lost some of the pleasant simplicity of the pure π-calculus. It might also not be evident at first sight what the inductive definition of the λ-π-hybrid would look like, due to the mixture of λ-names and π-variables in a single expression. Certainly, some problems remain, but having concentrated in this paper on the defence of using process algebras for the description of multi-modal discourse, we defer these matters to a more theoretical paper on the λ-π-hybrid.

Acknowledgements

Thanks go to Florian Hahn and Insa Lawler for work on gesture during the last years. Support by the German Research Foundation in the CRC 573, "Alignment in Communication", Bielefeld University, is also gratefully acknowledged.

References

Katya Alahverdzhieva and Alex Lascarides. 2010. Analysing language and co-verbal gesture in constraint-based grammars. In Stefan Müller, editor, *Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, pages 5–25, Paris.

Haskell B. Curry, Robert Feys, and William Craig. 1974. *Combinatory Logic*, volume 1 of *Combinatory Logic*. North-Holland Publishing Company, 3 edition.

Florian Hahn and Hannes Rieser. 2012. Non-compositional gestures. In *International Workshop on Formal and Computational Approaches to Multimodal Communication held under the auspices of ESSLLI 2012*, Opole.

Charles A. R. Hoare. 1985. *Communicating sequential processes*. Prentice-Hall International series in computer science. Prentice-Hall, Englewood Cliffs, NJ, 6. print edition.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 206–216, London, UK. Association for Computational Linguistics.

Adam Kendon. 2004. *Gesture – Visible Action as Utterance*. Cambridge University Press, Cambridge, NY.

Alex Lascarides and Matthew Stone. 2006. Formal semantics of iconic gesture. In David Schlangen and Raquel Fernández, editors, *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, Brandial'06, pages 64–71, Potsdam. Universitätsverlag Potsdam.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: the bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Andy Lücking, Thies Pfeiffer, and Hannes Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79.

Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter Mouton, Germany.

David McNeill. 1992. *Hand and Mind. What Gestures Reveal About Thought*. The University of Chicago Press.

Robin Milner. 1999. *Communicating and Mobile Systems: The π Calculus*. Cambridge University Press, Cambridge.

J. Parrow. 2001. An introduction to the π-calculus. In A. Ponse J.A. Bergstra and S.A. Smolka, editors, *Handbook of Process Algebra.*, pages 479–545. Elsevier.

Hannes Rieser. 2013. Speech-gesture interfaces. an overview. In *Proceedings of the 35th Annual Conference of the German Linguistic Society (DGfS)*, pages 282–283.

Insa Röpke. 2011. Watching the growth point grow. In *Proceedings of the Second Conference on Gesture and Speech in Interaction (GESPIN) 2011*.

D. Sangiorgi and D. Walker. 2001. *The π-calculus. A Theory of Mobile Processes.* Cambridge University Press, Cambridge.