On the primacy of language in multimodal communication

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
Nijmegen, the Netherlands
+31 24 3521541
janpeter.deruiter@mpi.nl

ABSTRACT

In this paper, I will argue that although the study of multimodal interaction offers exciting new prospects for Human Computer Interaction and human-human communication research, language is the primary form of communication, even in multimodal systems. I will support this claim with theoretical and empirical arguments, mainly drawn from human-human communication research, and will discuss the implications for multimodal communication research and Human-Computer Interaction.

Keywords
Multimodality, language, communication, interfaces.

INTRODUCTION

In an influential article on multimodal interaction, Oviatt [11] discusses and rejects ten common myths about multimodal interaction. The fourth myth is that *speech is the primary input mode in any multimodal system that includes it* [11, p.77]. I will defend the view that this is not a myth, but rather a deep truth, which multimodal researchers should be aware of, both in Human Computer Interaction (HCI) and in human-human communication research. In what follows, I will defend the *Linguistic Primacy Hypothesis* (LPH). I will formulate the LPH as a generalization of Oviatt's [11] formulation mentioned above, namely: "*Language is the primary input mode in any multimodal system that includes it*". By "language", I mean any modality (or to be more precise, *semiotic channel*, as defined in De Ruiter et al. [6b]) that uses a) arbitrary symbols with conventional meaning (lexical elements), and b) morphosyntactic rules that govern the combination of those lexical elements into larger utterances. In other words, speech, written or typed language, and the sign language of the deaf are all considered to be members of the category language, but for instance speech accompanying gesture is not.

I will defend the LPH by presenting a number of theoretical and empirical arguments to support it. Finally, I will discuss some of the implications of the LPH for multimodal communication research and multimodal HCI.

ARGUMENTS AGAINST THE LINGUISTIC PRIMACY HYPOTHESIS

This is not the first time in history that the truth of the LPH is questioned. In the late 1970ies, a number of communication researchers have claimed that nonverbal communication is far more important than language. For instance, Archer & Akert [1], asking their subjects to answer multiple choice questions about video fragments and transcripts, stated that "In fact, the current study provides no indication that verbal transcripts of interactions provide any independent contribution to accurate interpretation".

The claim that communication is mainly determined by nonverbal channels is analogous to the urban myth that we lose 90% of our body heat through our head. If that were actually true, one could safely go skiing naked, dressed only in a warm hat. In fact, it is only true that we lose 90% of our body heat through our head *if* we cover the rest of our body with insulating clothes. The relevance of this analogy becomes clear after realizing that [1] carefully removed verbal expressions from their materials that could have been informative, because they "did not want a simple test of audition". In other words, in their study, language did not get a fair chance.

As Brown [4] persuasively argued, it turned out to be the case in this and similar studies that nonverbal communication was predominant only *in the absence* of relevant linguistic information. When language was included, linguistic content turned out to be the best predictor of subjects' judgments of the emotional quality of the communication [8].

While the studies mentioned above focused mainly on the perceived emotional quality of the communication, more recent studies that have inspired multimodal researchers such as [10], have focused more on the *representational* aspects of communication.

It is obvious that communicating analog information such as spatial configurations can be cumbersome and inefficient in language, and that this is often done more efficiently using analog modalities such as gesture. However, for a fair comparison between language and non-linguistic modalities, it is important to also be aware of the communicative functions that language *can* perform, and the non-linguistic modalities *cannot.*

THE POWER OF LANGUAGE

Language can encode and transmit complex information that is very hard, if not impossible, to express in non-linguistic modalities. Some illustrative examples are logical connectives, such as conditionals, and temporal information, such as past and future. Imagine having to express the following simple sentences without using some form of language:

(1) If we don’t go now, we’ll miss the train.

(2) Last year I finally finished my book.

(3) Although it rains, I will go for a walk.

These examples are by no means an exhaustive demonstration of the expressive powers of language. Anyone who has ever played the family game “pictionary” will realize how hard it is to express certain ideas without resorting to the use of language. A picture may be worth a thousand words, but words are priceless. Oviatt [11] is of course correct in observing that gesture or other ‘analog’ channels might contain information that can only be expressed in language very inefficiently; my point here is that the reverse, expressing linguistic information in a non-linguistic modality is much harder, often even impossible.

MULTIMODAL FUSION

A strong argument for multimodal input processing is what is generally referred to as multimodal *fusion*. By combining information coming from different modalities it is possible to improve recognition quality and/or confidence. However, multimodal fusion relies fundamentally on different modalities containing redundant information. Since lip movements correlate with speech, they can in principle be used to improve speech recognition. However, many examples of multimodality in human-human communication show the use of what Engle [7] has termed *composite signals*. The information from gesture and the information from speech provide different aspects of a message. For composite signals to work properly, *both* modalities need to be reliable, and because the different components of the composite signal are by definition not correlated at the signal level, multimodal fusion will not improve their respective recognition accuracies. It is important to distinguish between fusion at the *signal* level and fusion at the *semantic* level. In the case of lip movements and speech, fusion is theoretically possible at the signal level, while in the famous “put that there” [3] example of deictic dereferencing, fusion is possible (and necessary) only at the semantic level. For semantic fusion to operate, both modalities need to have their own independent level of accuracy and confidence. In multimodal fusion, we cannot have our cake and eat it at the same time.

In fact, many existing implementations of both signal level and semantic level fusion provide evidence for the LPH because they crucially involve at least one linguistic modality. In many cases, most notably in the case of composite signals involving so-called *iconic* gestures [10], the gestures are generally not even interpretable without access to the affiliated speech [10, 5]. It appears that this often holds for other visual modalities such as facial expression as well [2].

NATURALNESS

Another strong and often quoted argument for multimodality is to improve the naturalness of the interaction. Just as humans use their face, eyes and hands to transmit messages to one another, machines could do so too, thereby more closely approximating face to face interaction between humans.

While this is a strong case for pursuing multimodal HCI applications, it is worth mentioning that the best way to make a multimodal interface appear *un*natural is by equipping it with slow and unreliable speech processing.

One of the main motivations to use Wizard of Oz (WoZ) technology in human factors experiments is that we suspect that leaving the speech modality to be processed by the machine will prevent us from obtaining interesting results. It is again the primacy of linguistic communication that is the reason for using WoZ procedures primarily for the linguistic modality.

EMPIRICAL EVIDENCE FROM HUMAN-HUMAN COMMUNICATION

First of all, it is truly amazing what humans can accomplish by using only the linguistic modality to communicate. Not only can we satisfy virtually every communicative need by using only speech (e.g. by telephone), but even email and chat, where we don’t even have access to paraverbal information such as prosody or voice quality, are highly effective in exchanging information, performing joint tasks, and maintaining social relationships.

In contrast, being in an environment in which we do not speak the language of our communicative partner will seriously hamper our communicative abilities, no matter how eloquently we gesture, draw pictures and faces, and pantomime. It is in these contexts that the lack of the previously mentioned capabilities of language become painfully obvious.

The SLOT experiment

In the COMIC project, we use an experimental paradigm called SLOT (Spatial Logistics Task, see [6b] for details). In this paradigm, two subjects are facing each other, both looking at their own copy of a “map” displayed on a graphical tablet front of them (see Figure 1).

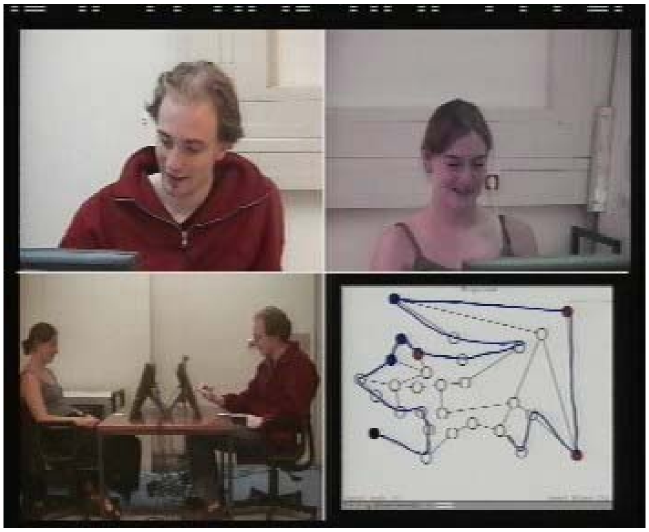

Figure 1. Snapshot from SLOT experiment

The subjects' task is to negotiate a route through the cities on the map while trying to minimize the "cost" of that route for themselves. In order to facilitate the negotiation process, subjects can draw on the map with an electronic pen. The tablet and electronic pen in SLOT implement a "shared whiteboard" metaphor.

One of the prime motivations for the development of SLOT is that we can selectively shut down certain modalities without changing the essential characteristics of the task. We can, for instance, put a screen or a one-way mirror between the subjects to block the transmission of facial expression and eye-gaze. Also, we can enable or disable the use of the electronic pen. During the pilot phase of SLOT, we also considered blocking speech (e.g. by letting the subjects wear headphones). However, even though the subjects could then still use the pen to draw suggested routes, the negotiation process crucially depends on exchanging, attacking, and defending *arguments* (motivations) for or against proposed routes. This is fundamentally impossible without speech, unless subjects use handwriting and write letters on the whiteboard to one another, which would defeat the purpose. We were mainly interested in the composite signals created by the parallel use of speech and pen gesture.

We ran a SLOT experiment with eight dyads that could use the pen, and eight dyads that could not. The latter group therefore had no choice but to describe proposed routes through the map using speech, whereas the former could (and did) draw them directly on the map. We expected that the total negotiation times would increase significantly for the dyads that could not use the pen. To our surprise, this was not the case at all. In Figure 2, the average negotiation times for the with-pen and without-pen conditions are shown.

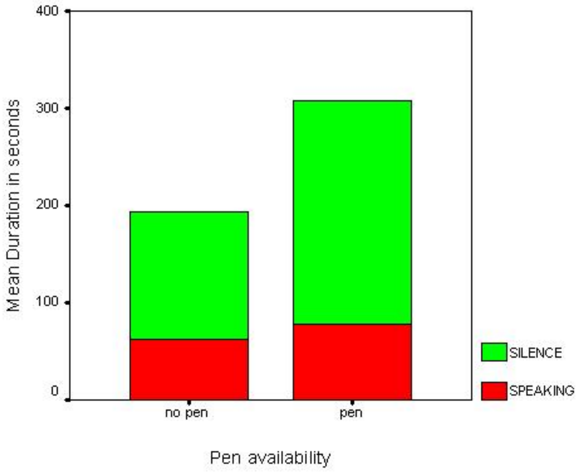

Figure 2. Average negotiation times

As can be seen from the graph, the without-pen dyads were even *faster* than the with-pen dyads, and there were no significant differences between the average speech durations. Importantly, the *quality* of the negotiated solution (measured as the sum of the incurred cost for the negotiated route for both subjects) was the same for both groups.

The main point to be made here is that the absence of the - in this context very natural and popular - pen modality did not lead to noticeable problems, neither in the efficiency of the negotiations nor in the quality of the negotiation outcome. Our subjects may have *preferred* to have used the pen, but they certainly didn't *need* it. Without speech, however, they could have drawn routes and perhaps used facial expression to display their evaluations of the routes, but they could not have *discussed* them.

IMPLICATIONS

So if indeed the LPH is correct, what are the consequences for multimodal communication research and HCI?

Let me emphasize that my arguments for the truth of the LPH are not in any way intended to discourage or discredit research efforts into multimodal communication, the use of multimodal fusion, or efforts to build maximally ergonomic, natural and efficient multimodal interfaces. On the contrary, I believe that an appreciation of the incredible flexibility and expressiveness of language can actually help us realize the goals of multimodal communication research.

First of all, by acknowledging the central role of language, we acknowledge the urgent need to improve language processing, especially at the input side. Speech recognition is often a serious bottleneck for the efficiency and naturalness of multimodal interfaces.

Second, as Levinson [9] has argued, speech is a very slow communication medium in terms of bits per second. The reason we can nevertheless communicate so efficiently in speech is that we can, in Levinson's words "piggyback meaning on top of meaning" [9, p.6]. In other words, not all relevant information needs to be contained in the signal. Verbal utterances are interpreted within a cognitive context. To model this remarkable human capacity in machines, it is necessary to interpret utterances against a background of contextually relevant knowledge. To model this functionality in machines, we need to have a) detailed, implementable knowledge about the implicatures, inferences and pragmatic conventions that are used by human language users, integrated with b) symbolic representations of the contextually relevant knowledge for the domain at hand. Multidisciplinary efforts involving both linguistics and Artificial Intelligence are essential for making our interfaces truly communicative.

Most importantly, for both human-human multimodal research and for multimodal systems it is essential that we develop annotation schemes and representational frameworks that enable us to represent the meaning of both linguistic and non-linguistic signals at the same representational level (see e.g. [6a]). This is especially challenging for those signals that do not carry representational meaning but are related to socio-emotional communication.

ACKNOWLEDGMENTS

I would like to thank Theo Vosse for his thoughtful comments on an earlier draft of this paper. Also, I am indebted to my colleagues in COMIC and in the Multimodal Interaction Project at MPI for many fruitful and inspiring discussions on multimodal communication.

REFERENCES

1. Archer, D. and R.M. Akert, *Words and everything else: Verbal and nonverbal cues to social interpretation.* Journal of Personality and Social Psychology, 1977. 35: p. 443-449.

2. Bavelas, J.B. and N. Chovil, *Visible acts of meaning - an integrated message model of language in face-to-face dialogue.* Journal of Language and Social Psychology, 2000. 19: p. 163-194.

3. Bolt, R.A., *Put that there: Voice and gesture at the graphics interface.* ACM Computer Graphics, 1980. 14(3): p. 262-270.

4. Brown, R., *Social Psychology*. 2nd ed. 1986, New York: The Free Press.

5. De Ruiter, J.P., *Gesture and Speech Production*. 1998, Doctoral Dissertation: University of Nijmegen.

6a. De Ruiter, J.P., 2003. A quantitative model of *Störung*. In: Kümmel, A. & Schüttpelz, E. (eds) *Signale der Störung*. Wilhelm Fink Verlag, München.

6b. De Ruiter, J.P., et al., *SLOT; a Research Platform for Investigating Multimodal Communication.* Behavior Research Methods, Instruments and Computers, 2003. 35(3): p. 408-419.

7. Engle, R.A. *Not channels but composite signals: Speech, gesture, diagrams, and object demonstrations in explanations of mechanical devices*. in *Twentieth Annual Conference of the Cognitive Science Society*. 1998. Madison, Wisconsin.

8. Krauss, R.M., et al., *Verbal, vocal, and visible factors in judgments of another's affect.* Journal of Personality and Social Psychology, 1981. 40: p. 312-319.

9. Levinson, S.C., *Presumptive Meanings; The Theory of Generalized Conversational Implicature*. 2000, Cambridge, Massachusetts: The MIT Press.

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

11. Oviatt, S., *Ten myths of multimodal interaction.* Communications of the ACM, 1999. 42(11): p. 75-81.