Physical, semantic and pragmatic levels for multimodal fusion and fission

Frédéric Landragin
LATTICE, 1, rue Maurice Arnoux, F-92120 Montrouge, France
Frederic.Landragin@linguist.jussieu.fr

1 Introduction

Multimodal fusion is linked to the integration of information in human-machine interactive systems where several communication modalities are proposed to the user. It is mainly used in input, for instance when the user can interact with the system using speech and 2D gestures. As a symmetric process, multimodal fission is linked to the repartition of information among several communication modalities. It is mainly used in output, and consists of one of the processes of an Intelligent Multi-Media Presentation System (IMMPS). The term 'fission' is sometimes used in input, for instance when the user's utterance has to be split into several requests (or speech acts or anything else). But this kind of segmentation is not multimodal and 'multimodal fission' keeps related to output.

Following state of the art studies and systems, especially in the field of natural language processing (in addition to the one of human-machine dialogue and multimodal interaction) [2, 3, 6, 8], multimodal fusion includes a first kind of fusion, at the level of the physical signal, and a second kind of fusion, at a semantic level. They can be called 'early fusion' and 'late fusion'. Temporal aspects like synchronicity between speech and gesture are sometimes included in the signal aspects, sometimes in the semantic aspects. Semantic fusion often corresponds to the integration of referential gestures to linguistic expressions. It is then closely related to the resolution of references to objects, in particular in dialogue where objects are visible for both the speaker and the addressee [5]. When gesture may be iconic, semantic fusion has also the role to unify the semantic features of speech and gesture, in order to reduce ambiguity and uncertainty. The term 'fusion' is sometimes used for the resolution of anaphora and ellipses, because some information from the linguistic history is 'fused' to the new utterance. Since it is not multimodal, we will ignore these aspects. As a statement, multimodal fusion usually groups the confrontation of temporal, prosodic, lexical, syntactic and semantic information in order to determine the global sense of the multimodal utterance.

Following works related to the design of IMMPS [1, 4, 8], the main role of multimodal fission is to determine which message will be generated within

each modality. When the global multimodal message has been determined by the dialogue manager, the monomodal parts of the message have to be sent to the speech synthesizer, to the graphics and animation manager, and, why not, to the haptic device. This process can be called 'information repartition' and occurs before the processes that are dedicated to the 'information rendering'.

In this paper we propose a unified vision of multimodal fusion and fission, where parameters for the first are exploited for the second and vice-versa. Following natural language processing (prosodic, lexical, syntactic, semantic and pragmatic analyses), we want to emphasize the pragmatic aspects that are not often studied in the context of 'symmetric multimodality' (see [8]).

2 Multimodal fusion

Instead of temporality and semantics, we prefer to distinguish several sub-processes for multimodal fusion: 'multimodal coordination', 'multimodal content fusion', and 'multimodal event fusion'.

- **Multimodal coordination**. This sub-process aims at associating two actions captured by different modalities, in order to construct a complete utterance. In the following we will only consider speech and gesture, but the principles are valid for each kind of multimodality. The inputs are the events constructed by the speech and gesture pre-interpretation modules. The output is a set of possible pairs of speech and gesture events. These pairs are hypothesis of fusionable events that will be processed by the 'multimodal content fusion' and 'multimodal event fusion' sub-processes. A trust level is associated with each pair, according to various parameters such as the type of gesture encapsulated in the events, the presence of deictic elements or the temporal alignment of the events. In particular, speech and gesture contain deictic events, like "this", "that" or a pointing on a graphical object, which are important clues to associate actions, but it is not compulsory. Temporal alignment is the other crucial element to efficiently perform the association of two actions.

- **Multimodal content fusion**. Two events that are coordinated are processed to obtain a coherent sense from incomplete information. This is the classical resolution of reference ("this object" + gesture, "put that there" + two gestures). This module has to take into account various cases. For example, a user may say "move that there" and produce a gesture characterized by a point of departure (that), a point of arrival (there), and a precise trajectory that reveals a precise way for the moving.

- **Multimodal event fusion**. Once multimodal contents are fused, the pragmatic forces of the monomodal acts have to be fused in order to produce a resulting complex act that contains all the pragmatic aspects provided by the user. General communicative act categories, which are independent from the communication modality or modalities, are inherited from classical natural language studies:

- "**Inform**" = the user informs the system, in a linguistic, gestural or multimodal manner. With a linguistic point of view, the classical form for such a speech act (also named 'say that' [7]) is the assertive form, e.g. "My destination is Paris.".
- "**Demand**" = the user requires the system to do something. The classical linguistic form is here the imperative form, e.g. "Give me the ways to go to Paris.". (the corresponding speech act category is 'tell to'). Concerning gesture, this is the case of all request oriented predefined forms, e.g., a cross-like gesture that means a deletion.
- "**Question**" = the user wants the system to give him an information. For a close question, the information is "yes" or "no". For an open one, it is a value, e.g., a duration with the example "How long does it take to go to Paris?". The classical linguistic form is the interrogative form, and an example of a gesture is a "?"-like trajectory.

The global category that subsumes the three previous ones is "**act**". This category is useful when it is difficult to label a gestural or linguistic message with a particular act. In this case, the "act" category is exploited in order to make the event fusion possible. In fact, the "act" category allows any fusion with demand, inform, or question categories. This mechanism is exploited for instance for simple pointing gestures. Concerning the events from gesture and speech that are compatible, some semantic aspects have to be emphasized. For example, a "demand" linguistic event (like "move that object") and a "demand" gestural event (like a "delete" meaning trace on the related object) are compatible events only if their semantic contents are compatible. Then, multimodal event fusion relies on multimodal content fusion.

3 Multimodal fission

Multimodal fission consists of splitting the information into several parts considering the presentation aims, means and context. Now, information can be split at different levels. At the signal level, the information, considering its nature, is sent to the correct communication channel. This is typically the case for a video, the sound track being sent to the auditory channel and the visual track to the visual channel. This is also the case for a linguistic utterance accompanied by one or more deictic gestures, such as "I am putting that there" with two gestures, one for "that" and one for "there". In this example, the IMMPS must be aware of the duration of the speech synthesis in order to provide the gestures, e.g., visual feedbacks, at the right moments. Splitting and synchronizing at the signal level is then a kind of multimodal fission, and is strongly linked to the constraint-based repartition over the communication channels.

At a semantic level, the information content can be dissociated over several modalities in order to better manage its complexity and to simplify the resulting monomodal messages. One important example related to human factors consists of displaying the part of the information that requires an important amount of

persistent attention, and of verbalizing the part whose only aim is to capture selective attention. Splitting at a semantic level appears as another kind of multimodal fission, which is linked to a preference-based repartition.

At a pragmatic level, the message illocutionary force can be dissociated over several modalities in order to simplify the illocutionary force of each monomodal message. For instance, an informative message that needs an acknowledgement of receipt from the user can be split into two messages: a first one that verbalizes the 'inform' and a second one that 'demands' the acknowledgement using a text box. To us, this is a third kind of fission, as important as the previous ones, although it has not been so studied in the literature. To conclude:

- At the physical level there is **multimodal coordination** for input signal processing and **multimedia coordination** for output processing;
- At a semantic level there is **content fusion** for input message processing and **content fission** for output message processing;
- At a pragmatic level there is **event fusion** for input event processing and **presentation act fission** for output event processing.

References

[1] Bernsen, N.O., 1996. A Reference Model for Output Information in Intelligent Multimedia Presentation Systems, in: *Proceedings of the ECAI'96 Workshop on Intelligent Multimedia Presentation Systems*, Budapest, Hungary.

[2] Bos, E., Huls, C., Claassen, W., 1994. EDWARD: Full Integration of Language and Action in a Multimodal User Interface. *International Journal of Human-Computer Studies* 40, pp. 473–495.

[3] Elting, C., Rapp, S., Möhler, G., Strube, M., 2003. Architecture and Implementation of Multimodal Plug and Play, in: *Proceedings of ICMI 2003.*

[4] Foster M.E., 2005. Interleaved Preparation and Output in the COMIC Fission Module, in: *Proceedings of ACL 2005 Workshop on Software*, Ann Arbor.

[5] Landragin, F., 2006. Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems. *Signal Processing* 86(12), pp. 3578–3595.

[6] Martin, J.-C., Buisine, S., Pitel, G., Bernsen, N.O., 2006. Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. *Signal Processing* 86(12), pp. 3596–3624.

[7] Sperber, D., Wilson, D., 1995. *Relevance. Communication and Cognition (second edition)*, Blackwell, Oxford UK and Cambridge USA.

[8] Wahlster, W., 2003. Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression, in: *Proceedings of the 26th German Conference on Artificial Intelligence.*