TR-IT-0013

Human-Oriented Design and Human-Machine-Human Interactions in Machine Interpretation

Christian Boitet

1993.8.30

This report first tries to demonstrate that future Dialogue Interpreting Telecommunications systems will necessarily integrate a human interpreter acting as supervisor and "warm body", rely on interactive disambiguation by the interlocutors, and be equipped with multimodal facilities. It then examines possible hardware configurations, and sketches a rough scenario for such a Human-Assisted Machine Interpretation system. Finally, direct and indirect "human-machine-human" (H-M-H) interactions are analyzed, in particular those occurring during interactive disambiguation. Related software engineering and lingware engineering research problems are briefly discussed, in particular that of managing parallel multimodal interactions.

This report is a synthesis and extension of two communications:

[1] Boitet, Ch. *Practical Speech Translation Systems will integrate human expertise, multimodal communication, and interactive disambiguation.* Proc. MTS-IV, Kobe, 18–22 July 1993, 173–176.

[2] Boitet, Ch. & Loken-Kim, K. H. *Human-Machine-Human Interactions in Interpreting Telecommunications.* Proc. International Symposium on Spoken Dialogue, Waseda University, Tokyo, 10–12 November 1993, to appear.

The research reported here was conducted while I was staying at ATR Interpreting Telephony Research Laboratories, and then at ATR Interpreting Telecommunications Research Laboratories, as visiting researcher from GETA, IMAG, UJF&CNRS, France. I would like to heartly thank ATR for its constant support and very favorable research environment; and its members, from President through supervisors through researchers through secretaries to security personal, for all the personal help which they extended to me in so many ways at so many occasions.

Interpreting Telecommunications Research Laboratories
2–2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Contents

Abstract

We argue that future Dialogue Interpreting Telecommunications systems will necessarily integrate a human interpreter acting as supervisor and "warm body", rely on interactive disambiguation by the interlocutors, and be equipped with multimodal facilities.

In this context, it will be necessary to manage a variety of multimodal interactions in parallel. *Direct* interactions will occur between the interlocutors, the machine, and the interpreter. *Indirect* interactions will allow the human participants to exert control on direct interactions between two other participants. Indirect interactions present new interesting aspects, notably the need for *progressive* translation.

We use the term "human-machine-human" (H-M-H) interactions to cover both direct and indirect interactions. We analyze them in some detail, and examine associated problems. One of them is managing parallel multimodal interactions, which suggests an object-oriented, distributed design. ATR is currently developing a multimodal simulator in this perspective.

Keywords:

H-M-H Interactions, Interactive Disambiguation, Interpreting Telecommunications, Machine Interpretation, Parallel Multimodal Interactions, Speech Translation.

Introduction

The term "translation", while more general than the term "interpretation", is preferably used to denote the translation of text, while "interpretation" denotes only the translation of speech. One further distinguishes between "consecutive" and "simultaneous" interpretation. Although the ultimate aim of research in Speech Translation (ST) is to produce Machine Interpretation (MI) systems simulating simultaneous interpretation, we don't have even the beginning of a model for it, so that consecutive interpretation is the current goal.

Current ST prototypes produce an intermediate written form of the original spoken message, because they work by sequentially combining speech recognition (SR) and machine translation (MT). Even if more integrated designs are implemented in the future, the tasks of speech recognition and natural language analysis (NLA) are so difficult that such a written form will remain very useful, for checking and editing purposes. In other words, MI systems should perhaps more appropriately be called ST systems. We will use both terms.

ATR has recently embarked on a seven year project on "Interpreting Telecommunications". Future Interpreting Telecommunications systems could be applied to situations with one user only[1], two users (dialogue), and several users (teleconference). This paper concentrates on the interpretation of dialogues.

[1] For example, dictating a commentary of a video scene with a view to generate subtitles in several languages, or preparing information messages during an international event for broadcasting in several languages.

A previous project, on "Interpreting Telephony" [10, 14, 15], has shown that, even in the context of a very restricted task and sublanguage, the system should at least provide feedback on what happens at the other end, if possible through various modalities[2]. Future telecommunication systems, whether used with translation or not, will be multimodal, and allow users to communicate directly by seeing each other's face and gestures, and by manipulating shared objects, such as maps, drawings or virtual objects. Hence, MI will be multimodal.

Let us call the machine M and the two interlocutors A and B. In MI, direct interactions between A and B will be non-verbal, but A and B will talk and listen to M. Due to the inherent limitations of fully automatic techniques, multimodal disambiguation dialogues (A-M, B-M) will also be necessary.

For practical systems handling generic tasks, a human interpreter (called X, for "expert") would have to be integrated in the overall design. X would supervise a certain number of dialogues ("sessions"), set system parameters for each of them, and act as a "warm body" if case of need. X would interact with M to get information on the dialogue between A and B, and perhaps to get on-line terminological help.

It will also be essential that each interlocutor be in control as well of his "interaction domain" as of some aspects of other users' domains. For example, A might stop the interactive disambiguation at the other end (B-M) if the current rough translation is enough for understanding, or X could interrupt interactions between A, B, and M to take over.

We propose the term "human-machine-human" interactions to cover the *direct* interactions (H-M and H-H) between any two of the four participants, as well as the *indirect* interactions of a human participant interfering with another interaction (e.g., the control of A over B-M or B-X).

The remaining of the report is organized as follows.

In the first part, we elaborate on the supervised and "Human-Aided Machine Interpretation" (HAMI) architecture sketched above for future practical systems: some possible applications; the necessity of introducing an expert interpreter; and the necessity of relying on user assistance. In the second part, we discuss multimodality and possible hardware configurations. In the third part, we analyze in more detail the H-M-H interactions to be handled: a rough scenario; multimodal interactions for disambiguation, and some aspects of handling parallel multimodal interactions.

---

[2] The video feed-back incorporated in the final ASURA demonstration [10, 15] was indeed a key element for its success.

I. Supervised and user-aided (HAMI) architecture

1. Possible applications

Applications envisaged for Machine Interpretation include assistance to professional or personal telephone dialogues, such as car rental, medical consultation, scheduling of meeting, greetings, explanation of itinerary...; teleconference; and multilingual dissemination of information. In this paper, we concentrate on dialogue situations.

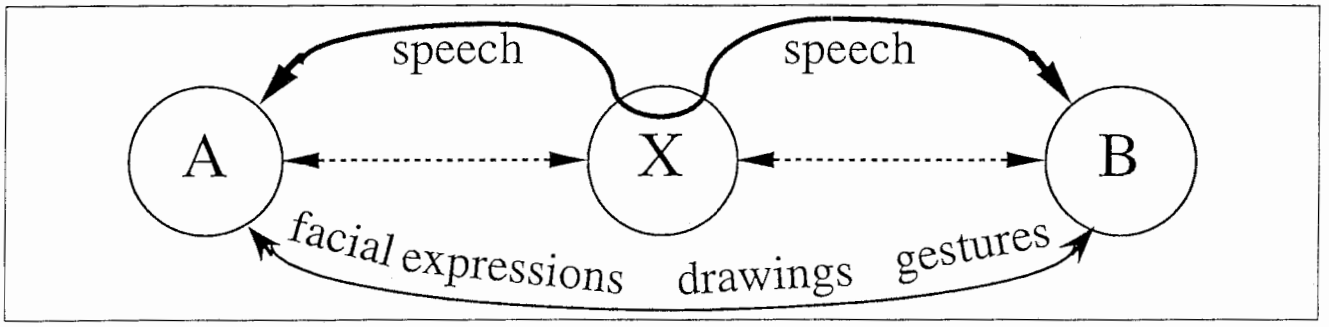


Figure 1: Interactions in human interpretation

Human interpretation is most often performed face to face, with the two speakers (A and B) and the interpreter (X) being in the same room. However, telephone interpretation services have long been available.

- KDD offers an interpretation service for transcontinental conversations. Here, the three participants are in three different locations. X begins by asking a few questions to A and B, and then translates the dialogue between them in consecutive mode. The session also ends with two short dialogues between X and each interlocutor.
- NTT offers an interpretation service to help foreign patients visiting Japanese doctors. A and B are in the same room and share the same telephone headset. X acts really as a broker: communication between A and B occurs through two dialogues, X-A and X-B, and not through a direct (translated) dialogue between A and B.
- AT&T offers the same kinds of services, and also a "patched interpretation" service. For example, if a foreign client calls Hertz because his car has broken down, Hertz can dial this service, and the interpreter is "patched in".

Figure 1 above shows the interactions in (human) consecutive or simultaneous interpretation. The thick line represents the main flow of communication, which is between A and B and goes through the human interpreter X. The thinner line stands for the direct communications between A and B, through facial expressions, drawings, gestures, or the manipulation of objects. The dotted, thinner lines represent clarification dialogues between the interpreter and the speakers.

2. Necessity of introducing an expert interpreter

How can such situations be automated? Replacing the human interpreter by a fully automatic system is certainly not going to work, because of *inherent* limitations in what can be expected of fully automatic speech recognition and language analysis in the foreseeable future and for practical settings.

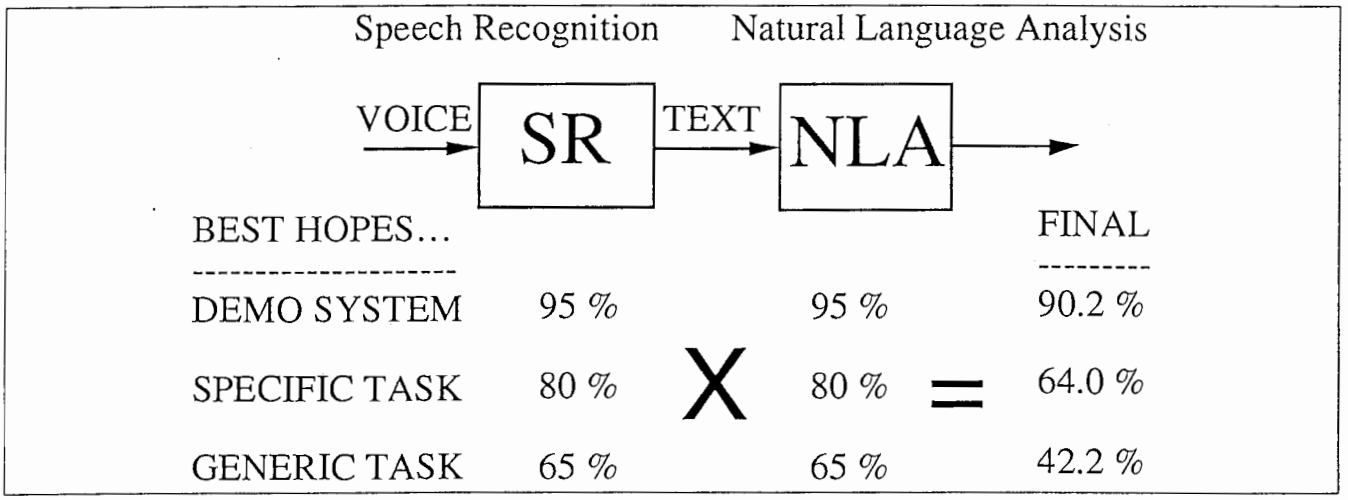


Figure 2: The "squaring" problem in Speech Translation

As other components contribute only marginally to the error rate, MI is faced with the "squaring" problem (figure 2).

If a demo system[3] handles 95%[4] of the utterances correctly, and similarly for the language analysis part, it will get 90.2% correctness, which is certainly impressive.

For a specific task[5], which might better be handled by a multilingual expert system than by a ST system, a success rate of about 80% is really the best hope for both components, but gives a combined success rate of only 64%.

The case of a generic task, with 3000/5000 words, speaker independence, spontaneous speech, and a large grammatical coverage, is almost desperate: the best hope would be 65% for each component[6], giving an overall success rate of only 42.2%, clearly unacceptable by human users.

Could the current technology, black-box, sequential, and speech-only, as diagrammed below, be applied to some realistic situations?

---

[3] About 700/1000 word forms, speaker adaptation, small finite set of possible sentences.

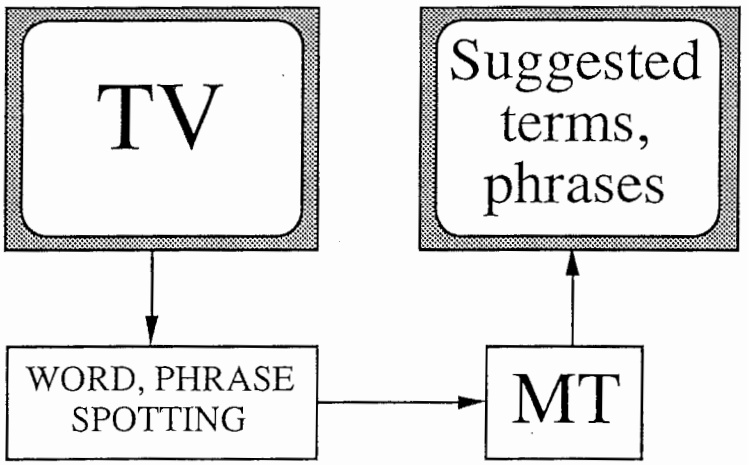
[4] These percentages, while based on figures found in the literature or presented at conferences, should be taken only as rough estimates.

[5] About 1500/2000 words, possible speaker adaptation, severely constrained language model.

[6] As a matter of fact, that is the best which commercial, large coverage MT systems can do. But these systems are used for *assimilation* of information in foreign languages, not for *dissemination*, and even less for human to human communication. Specialized systems reach 80%, which makes them cost-effective for use by professional post-editors producing high-quality final output. Extremely restricted MT systems such as METEO [6] reach indeed 95-97%.

This architecture can not be used in interpreting telecommunications...

Yes, and in particular to situations where a foreigner tries to get some understanding of an ongoing discourse, for learning purposes, or simply for getting some information, as in the case of TV news iluustrated below.



...but could be applied to other interesting tasks.

However, that kind of application is far away from the goals of Machine Interpretation, and we have to return to our problem of raising the overall success rate to the high level required by users int the context of interpreting telecommunications.

Integrating an expert interpreter X ("warm body") seems unavoidable if the overall success rate is not extremely high (perhaps more than 90%). It is also advisable for the same reasons which lead telephone companies to keep manual operators in parallel with automated systems.

Hence, we have to settle for, at least, a Human Aided Machine Translation (HAMT) architecture, where the human interpreter X is not only a "warm body", but also an expert of the system: X should supervise the system, advise its users, and adjust its parameters and/or take over if necessary.

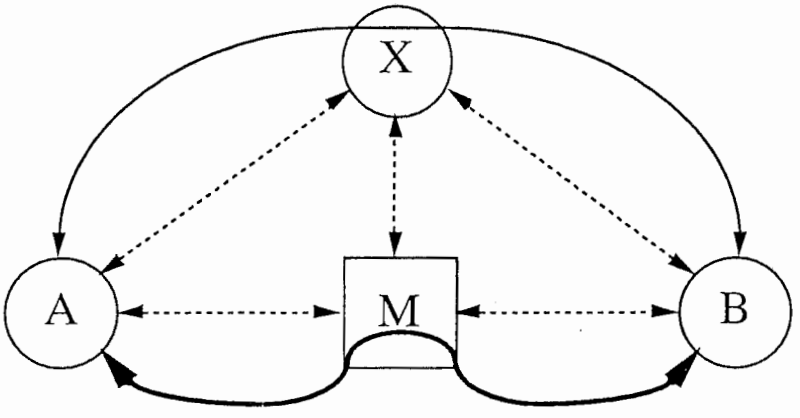


Figure 3: Human Aided Machine Translation (of one dialogue)

On the other hand, if A and B must rely on X for more than a relatively small fraction of the utterances, say 15%, they will tend to rely always on X and not to use M in the first place. Then, we must find a way to raise the overall success rate from about 64% in the case of a specific task, or from a dismal 42% for a generic task, to at least 85%. If we succeed, the interpreter will be able to attend to several conversations in parallel, which would really bring about a real progress in "interpretation productivity".

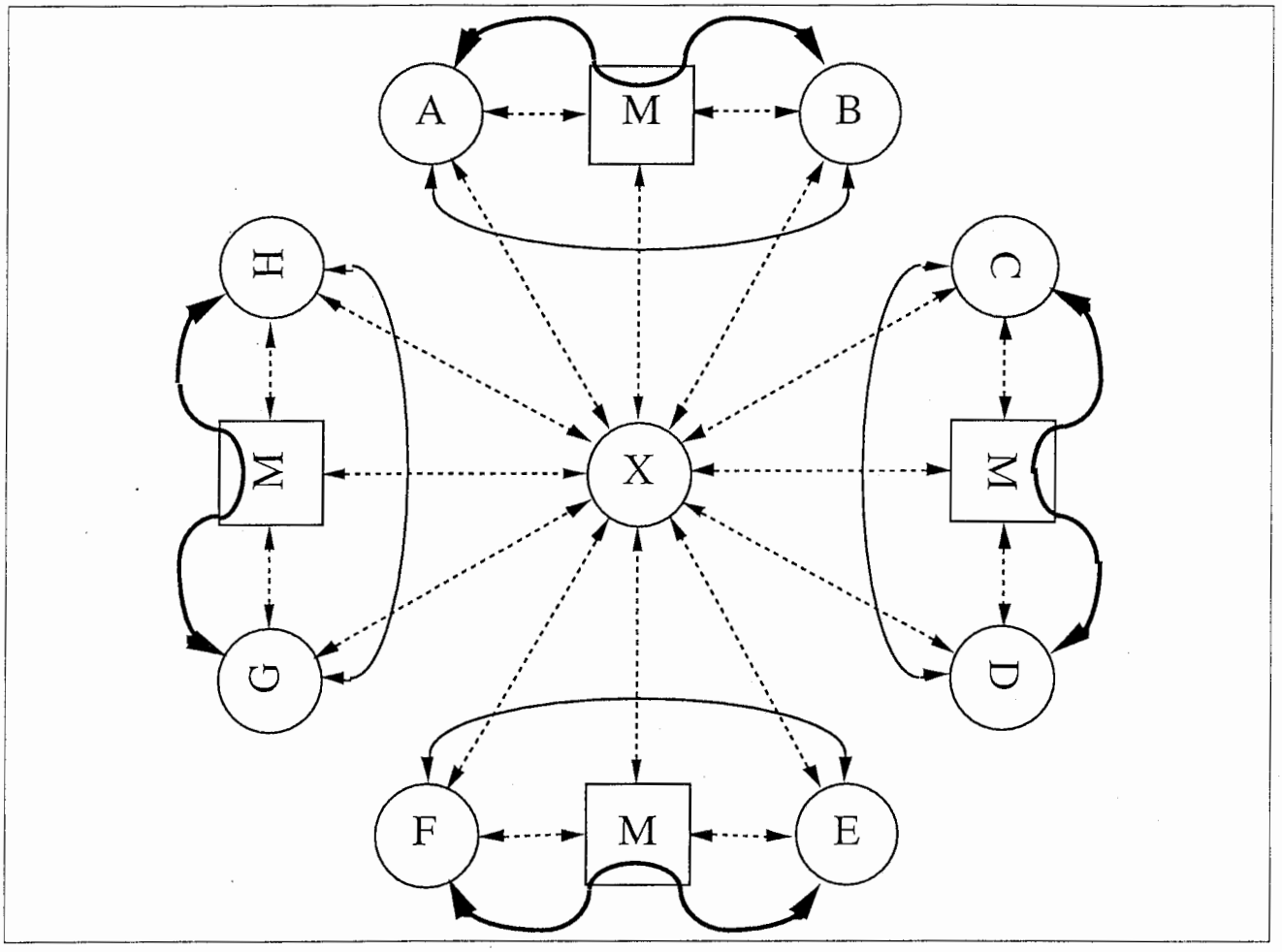


Figure 4: The interpreter as supervisor in a HAMT system

How, then, could we improve the overall success rate in "automatic mode"?

3. Necessity of relying on user assistance

Improving the user-friendliness of the system, while always desirable [16], would not help raising the success rate. However, it could perhaps lower the goal, to, say, 75%. For example, it should be possible to tune parameters controlling the perception of other agents (interlocutor, system, interpreter...), to monitor the progression of the translation process, and even to interrupt it (because the meaning has already been understood, or in order to correct the previous utterance), thereby reducing waiting time and associated frustration.

This last point suggests the interesting possiblity of building "progressive" MI systems, which would output successive states of the translation, on appropriate media, beginning with isolated words, then phrases, then complete raw translation, to finish with polished translation if available.

Equipping the system with knowledge about the generic task, i.e. with a partial ontology, is also a possibility, but is likely to be an overkill, because the speakers will in any case be far better at understanding the task at hand. Anyway, even with a complete ontology, an automatic system, as a human being for that matter, will never be able to fully disambiguate even clean, typed sentences, because part of the necessary information depends on the pragmatic, intentional context, and is accessible only by asking[7].

The only viable approach, then, seems to abandon the ideal goal of fully automatic, high quality speech recognition and natural language analysis components functioning as black boxes, and to rely on the assistance of the interlocutors themselves, using a Human-Aided Machine Interpretation (HAMI) architecture. Then, one expert interpreter could supervise and service several dialogues at the same time.

II. Multimodality and usability

1. Multimodal configurations

Experimental studies [20] have confirmed that human to human communication is far more effective if several modalities can be used. For instance, pointing at a map to explain an itinerary is far better than to explain only with words. Or, if the interlocutor does not catch a proper name after repeating it and perhaps spelling it, writing it on a keyboard or a notepad solves the problem immediately.

Hence, multimodality will be expected by users of communication systems, with or without human or automatic interpretation. Multimodality will also be required to make interactive disambiguation possible, and in particular for allowing users to guide the systems while talking.

The most simple telephone sets today have a 12-key keypad, which could be used for simple interactions, but probably not for interactive disambiguation. However, future hardware configurations will offer enough possibilities, even for applications concerning the *general public.*

Phones equipped with videotext terminals, such as the French "minitel" or touchscreen terminals, are already widely used. In the near future, general users will connect their notebook computers, or more specialized "personal communicators", to normal telephone outlets, and interact by speaking, writing text, pointing at maps, and drawing diagrams.

The office microcomputers of *professional users* have already graphics and sound, with light pens, graphic tablets, scanner and video as options. At ATR, a multimodal prototype simulator based on two NeXT stations equipped with all these options is currently being built under the direction of the second author. Finally, virtual reality will soon be available to *executives* participating in teleconferences.

---

[7] For these very reasons, interactive disambiguation through the "augmentor" had to be introduced in KBMT-89 [3].

Asymmetric situations, where the partners don't have the same equipment, may be quite frequent. We propose to define abstract objects, such as a "common working domain", or a "current partial analysis", and an object-oriented architecture, to ensure that the system will present them in the most appropriate ways on each configuration.

2. Rough scenario

Relative to each A-B "session", the HAMI system may be in one of two major modes, "manual" and "automatic", depending on whether X is acting as interpreter or not.

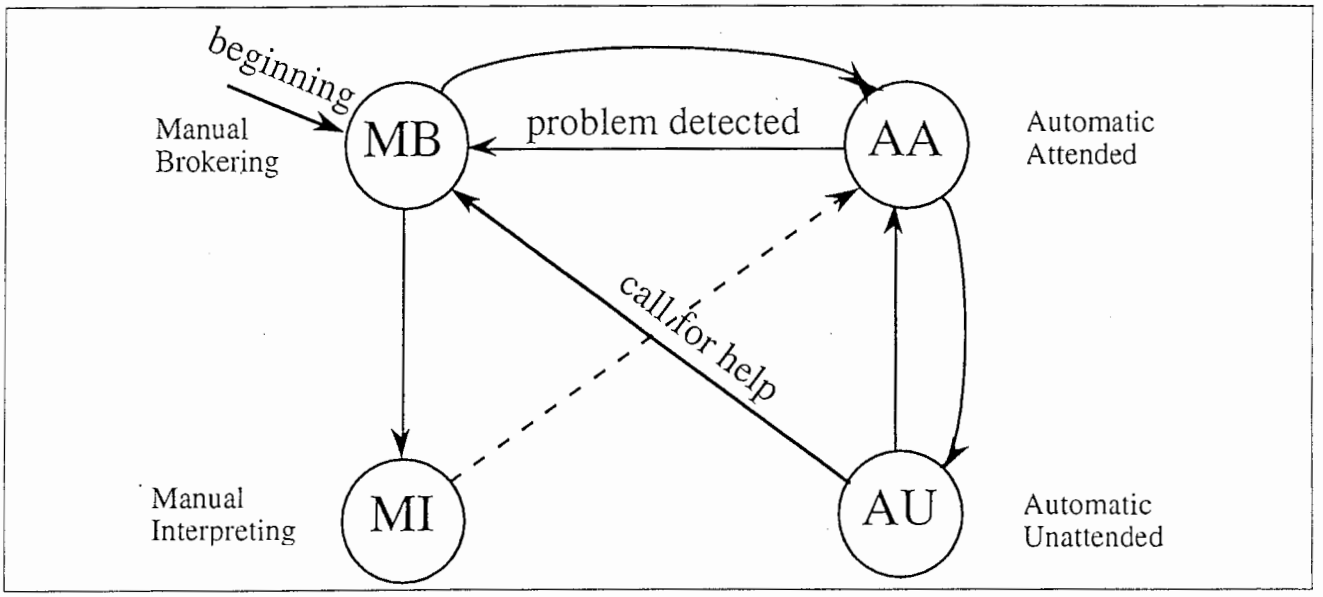
The manual mode has two minor modes.

- In "manual brokering" mode, X directs the flow of communication, by conducting two dialogues, X-A and X-B. This happens in particular at the beginning of the A-B session, where X sets parameters, such as task, proper names, or social relationship, used by the system to constrain its search space. It may also happen in cases such as the doctor-patient situation.
- In "manual interpreting" mode, X interprets the dialogue between A and B in the usual fashion.

The automatic mode has also two minor modes.

- In "automatic attended" mode, X follows the A-B session (and perhaps others) and resets system parameters if the success rate degrades because A and B don't guide the system well enough.
- In "automatic unattended" mode, X is not watching the A-B session, so that A or B must call for help if some difficulty arises, through any appropriate channel, e.g. by hitting "###" on the phone keypad.

After beginning in manual brokering mode, X would put the A-B session in automatic attended mode for a while. If A and B seem to do well, X would attend to another task and put A-B in automatic unattended mode. From time to time, X would put A-B back in automatic attended mode, and perhaps in manual brokering mode to give some short but useful advice to A and B on how to best use the system.



Mode sequence in rough scenario

If called, X would put A-B in manual brokering mode, ask A and B about their difficulties, and decide either to put them back in automatic mode, or to continue in manual interpreting mode[8]. In principle, one could to go back from that mode to automatic mode, but users are likely to want to finish their whole session in company of the "warm body"!

III. parallel multimodal interactions

1. Multimodal interactions for disambiguation

Voice disambiguation gives rise to two interactions. First, the SR component should indicate (visually or acoustically) its level of difficulty of recognition, if possible with an appropriate diagnostic (e.g. too fast, too slow, microphone not in place...). Second, the speaker could "clean" the input, by editing fragments of the intermediate written form marked as doubtful by SR, or by repeating them aloud.

On the linguistic side, users will be encouraged to *guide the system by "active" multimodal disambiguation.* Examples are: pressing a button to indicate the end of a sentence within a speech period; navigating through a graphic representation of the task domain in order to dynamically restrict the expected vocabulary; indicating the communicative type of the current utterance (assertion, question, request, advice...) to facilitate semantic and pragmatic interpretation; pointing at an icon or at a map to clarify an anaphoric reference; and editing the intermediate written form, e.g. by inserting or correcting punctuation marks.

In general, these interactions occur in parallel with the speech interaction.

Previous studies [16] report that up to 30% of utterances in bilingual telephone conversations using a human interpreter concern the clarification of what the speaker says, so that to have the system ask questions is a practical possibility.

Interactions related to that *passive, or system-initiated, disambiguation* process will also be multimodal. *We don't however expect users to frequently use two modalities in parallel at a given moment.*

Questions could be asked in various ways [1, 3-5, 7, 11, 13, 16, 17, 22]: for example, as items to select in menus; as written or spoken yes-no questions; or as intuitive graphic representations of the system's best guesses, to be corrected if necessary. Users should be allowed to set their preferences.

Questions may concern all kinds of ambiguities, lexical, syntactic, semantic, pragmatic, and communicative. It is a difficult problem to present ambiguity problems in intuitive ways, so that no specialized knowledge would be required of users. We plan to use the above-mentioned simulator to experiment with various techniques.

---

[8] In this case, X might transfer the communication to another interpreter, perhaps not expert in the use of the HAMI system.

2. Some aspects of handling parallel multimodal interactionss

Parallelism of interactions poses several problems. On the software side, concurrent accesses to the same object, e.g., to the same map, must be handled properly. Also, two participants (e.g., A and X) might want to stop an interaction at the same time (e.g., the B-M dialogue), which would give rise to a three-way conflict. This calls for a distributed, object-oriented system architecture.

On the ergonomy side, there should be a clearly defined communication discipline, for each mode of the system. For example, in manual brokering mode, A should not really interrupt B-X, but only warn B and X of his desire to come in. This discipline should be known and enforced by the system.

On the lingware side, the existence of parallel interactions offers new possibilities. For example, if a doctor says "ashi" and points at a foot on an anatomic diagram, the system should not ask whether "foot" or "leg" is meant. On the other hand, this calls for new "multimodal" linguistic formalisms, on which research is just beginning.

Also, some of the planned interactions depend on the availability of incremental, or "progressive" translation. Although there has been some previous research on incremental language generation, SR and NLA components should be adapted to deliver their "current best" solution, without slowing them down too much, or completely redesigned in this perspective.

Conclusion

This report has shown that future MI systems will give rise to a great variety of multimodal H-M-H interactions, analyzed them in some detail, and outlined certain interesting associated research problems. I hope that experiments with the simulator currently under construction at ATR will allow us to refine these ideas in the future, and to progress in the study of these problems.

Acknowledgments

Thanks are due to Dr. Y. Yamazaki, President of ATR-ITL, and T. Morimoto, Head of Department 4, for their support and encouragement; K. H. Loken-Kim, for numerous discussions of these ideas and for launching the multimodal simulator project; F. Yato, for his contribution to the simulator, and M. Seligman, for many stimulating discussions and stylistic remarks.

-o-o-o-o-o-o-o-o-o-o-

Bibliography

[1] **Blanchon H. (1992)** A Solution to the Problem of Interactive Disambiguation. Proc. COLING-92, 1233-1238.

[2] **Boitet C. (1988)** *Software and lingware engineering in modern M(A)T systems.* In "Handbook for Machine Translation", Bátori, ed., Niemeyer.

[3] **Boitet C. (1989)** *Speech Synthesis and Dialogue Based Machine Translation.* Proc. ATR Symp. on Basic Research for Telephone Interpretation, Kyoto, December 1989, 6-5-1—22.

[4] **Brown R. D. (1989)** *Augmentation.* Machine Translation, 4, 1299-1347.

[5] **Brown R. D. & Nirenburg S. (1990)** *Human-Computer Interaction for Semantic Disambiguation.* Proc. COLING-90, Helsinki, vol. 3/3, 42-47.

[6] **Chandioux J. (1988)** *10 ans de METEO.* In "Actes du séminaire international sur la TAO", A. Abbou, ed., OFIL, Paris, mars 1988, 169—173.

[7] **Chandler B., Holden N., Horsfall H., Pollard E. & McGee Wood M. (1987)** *N-tran Final Report.* Alvey Project, 87/9, CCL/UMIST, Manchester.

[8] **Hutchins W. J. (1986)** *Machine Translation: Past, Present, Future.* E. Horwood, ed., John Wiley & Sons, 382 p.

[9] **Hutchins W. J. & Somers H. L. (1992)** An Introduction to Machine Translation. Academic Press, 362 p.

[10] **Kurematsu A., Yato F., Morimoto T. & Yamazaki Y. (1993)** *Important Issues for Automatic Interpreting Telephone Service.* Proc. International Symposium on Spoken Dialogue, 10-12 Nov. 93.

[11] **Maruyama H., Watanabe H. & Ogino S. (1990)** *An Interactive Japanese Parser for Machine Translation.* Proc. COLING-90, Helsinki, vol. 2/3, 257-262.

[12] **Melby A. K. (1981)** *Translators and Machines - Can they cooperate ?* META, **26/1**, 23-34.

[13] **Melby A. K., Smith M. R. & Peterson J. (1980)** *ITS : An Interactive Translation System.* Proc. COLING-80, Tokyo, 30/9-4/10/80, M. Nagao, ed., 424—429.

[14] **Morimoto T., Suzuki M., Takezawa T., Kikui G.-I., Nagata M. & Tomokiyo M. (1992)** *A Spoken Language Translation System: SL-TRANS2.* Proc. COLING-92, Nantes, juillet 1992, C. Boitet, ed., ACL, vol. 3/4, 1048—1052.

[15] **Morimoto T., Takezawa T., Yato F., Sagayama S., Tashiro T., Nagata M. & al. (1993)** *ATR's Speech Translation System: ASURA.* Proc. EuroSpeech'93, Berlin, 21-23/9/83, 4 p.

[16] **Neal J. G. & Shapiro S. C. (1991)** *Intelligent Multimedia Interface Technology.* In "Intelligent User Interfaces", ACM Press & Addison-Wesley, New-York, 11—44.

[17] **Nirenburg S. & al. (1989)** *KBMT-89 Project Report.* Center for Machine Translation, Carnegie Mellon University, Pittsburg, April 1989, 286 p.

[18] **Nyberg E. H. & Mitamura T. (1992)** *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains.* Proc. COLING-92, Nantes, 23-28 July 92, C. Boitet, ed., ACL, vol. 3/4, 1069—1073.

[19] **Oviatt S. L. (1993)** *Toward multimodal support for interpreted telephone dialogues.* In "Structure of Multimodal Dialogue", M. M. Taylor, F. Néel & D. G. Bouwhuis, ed., Elsevier, Amsterdam, in press.

[20] **Oviatt S. L. & Cohen P. R. (1991)** *Discourse structure and performance efficiency in interactive and noninteractive spoken modalities.* Comp. Speech & Lang., **5/4**, 297—326.

[21] **Sullivan J. W. & Tyler S. W., ed. (1991)** *Intelligent User Interfaces.* ACM Press, Addison-Wesley, New-York, 472 p.

[22] **Tomita M. (1986)** *Sentence Disambiguation by Asking.* Computers and Translation, **1/1**, 39-51.

-o-o-o-o-o-o-o-o-o-o-