TR-IT-0070

Specification of first EMMI Experiments for the MIDDIM project in the context of Interpreting Telecommunications

Georges FAFIOTTE & Christian BOITET

1994.9.7

This note specifies the first EMMI-based experiments conducted during our research stay at ATR, in the framework of the MIDDIM project.

In these experiments, we want to test what influence variations in speech recognition ability and language proficiency of the interpreter, and in domain knowledge of all participants, will have on the number of disambiguation questions asked by the (automatic) interpreter.

We expect that the records, videos, and transcripts resulting from those experiments will provide for the refinement of the preliminary user modeling we propose here.

We begin by specifying the envisaged dialogue scenario, and the most relevant characteristics of the participants. Then we sketch a methodology for EMMI-based simulation situations. Finally, we propose a first set of experiments.

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

Table of Contents

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

Specification of first EMMI Experiments for the MIDDIM project in the context of Interpreting Telecommunications

Georges FAFIOTTE & Christian BOITET

GETA, IMAG (UJF & CNRS
BP 53, F-38041 Grenoble Cedex 9

research stay at

ATR Interpreting Telecommunications
2-2 Hikari-dai, Seika-cho, Kyoto 619-02

Wednesday 24 August 1994

Introduction

In this note, we want to specify the first EMMI-based experiments we would like to conduct while working here at ATR, in the framework of the MIDDIM project.

Recall the aim of this project is to study multimodal disambiguation techniques in the two contexts of Interpreting Telecommunications of spoken dialogues, and of Dialogue-Based Machine Translation of written texts. In this note, we concern ourselves with the first context only.

In these experiments, we want to test what influence variations in speech recognition ability and language proficiency of the interpreter, and in domain knowledge of all participants, will have on the number of disambiguation questions asked by the (automatic) interpreter.

We expect that the records, videos, and transcripts resulting from those experiments will provide for the refinement of the preliminary user modeling we propose here.

We begin by specifying the envisaged dialogue scenario, and the most relevant characteristics of the participants. Then we sketch a methodology for EMMI-based simulation situations. Finally, we propose a first set of experiments.

1. Dialogue participants and their characteristics

1.1 Dialogue participants

We envisage here dialogues conducted in a multimodal situation, as opposed to speech only dialogues, although both varieties are supported by EMMI.

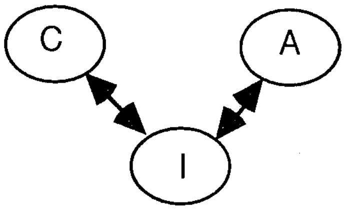
Subjects in the experiments are:

C, a Client,

A, an Agent,

I, an Interpreter,

the main situation being as depicted in the following diagram:



In the envisaged experiments, the automatic interpreter I will be replaced by a human "Wizard of Oz".

1.2 Characteristics of participants

We characterize the participants by their level of expertise, with respect to 3 main factors, namely:

- Linguistic Proficiency (Language Processing Ability) in the languages involved,
- Domain Knowledge,
- Auditory Aptitude (Speech Recognition Ability): in this context, we mean the ability to identify (i.e. to segment and recognize) words and phrases, even without understanding their meaning.

Other human factors, such as the participants' skill in manipulating the software environment, the interpreter's expertise in handling complex interpreting situations, the client's irrelevant interventionism, or the client's pertinent desire to constantly assess what's going on, could also be taken into account in the future.

1.3 Proposal for a scale for the three main factors

We may evaluate the characteristics intended for the target situation (machine-based Interpreting Telecommunications), and infer the Interpreter's abilities in our simulation situation.

Let's assume attributes coding to be as follows:

0	*none*
1	*poor, low*
2	*mediocre*
3	*average*
4	*fair*
5	*good*
6	*very good, expert*

A complete scope of values is not necessary in the context of the intended EMMI-based experiments, pertaining to realistic IT situations.

- For linguistic proficiency, we consider only: *none, low, average, fair, good.*

  In IT situations, Client and Agent should be rated *none* or *poor* in their partner's language, since bilingual users are not expected to use IT systems.

  The language processing ability of the machine, and therefore the linguistic proficiency of the Interpreter playing its role, is (optimistically?) expected to rank among: *low, average.*
- Proposed scores for domain knowledge are to be: *poor, fair, good.* Client and Agent will be credited with *fair* or *good.* The machine (Interpreter) domain knowledge will be first restricted to: *poor* to *average.*
- Auditory aptitudes (Speech Recognition ability) are suggested to be: *average, fair, good* or *very good,* with *average* to *very good* for the 'real' human actors (Client, Agent), and *average* or *fair* for the machine (Interpreter).

If we include in the future the factor of human familiarity with the environmental interface (for the Client and Agent only), it could be set as: *low, mediocre* or *average, good.* In the first experiments to be conducted, we consider them to be *average, good* for the Client and *good* for the Agent, who can be expected to use the system regularly. Other levels of skill could be studied later.

1.4 Setting up experiment factors.

We thus propose to focus on pertinent user profiles, i.e. relevant to the simulation situation and most likely to occur in later real interpreting telecommunications.

Situations are summarized in the following 'Subject/Factor Profile Table', where x–y indicates an interval, < x > an element, < x , y > a pair, and < x–x' , y–y' > all pairs < x" ,y " > with x" in x–x' and y" in y–y".

\ Factor **Subject **	**Language Proficiency** <L_C range, L_A range>	**Domain Knowledge**	**Auditory Aptitude** <L_C perc., L_A perc.>
Client C	< 4–6 , 0–1>	< 3–6 >	< 5–6 , 0 >
Agent A	< 0–1 , 4–6 >	< 3–6 >	< 0 , 5–6 >
Interpreter I (simulating a machine)	< 2–4 , 2–4 >	< 1–2 >	< 3–4 , 3–4 >
Total: **1,492,992** !!!	12*12*9=**1296** situat.	4*4*2 = **32** situat.	2*2*9 = **36** situat.

The total number of possible experiment patterns is obviously far too big, so we set up to reduce it.

Comments:

- For a meaningful experiment on interactive disambiguation, the human interpreter, playing the role of a machine interpreter, should not have a very high level of linguistic nor auditory ability.

- Interpreter's variation in auditory aptitudes may be caused artificially by using a vocoder and adjusting parameters.

- A column (and further experiments), could be added for a System Interface or Environment Mastery factor, with the following values:

- Client: < 0–3 > or at least < 0–2 >
  These marks account for 'naive' users. It is expected that this factor might also have some effect on the frequency of disambiguation sequences.
- Agent: < 3 > (since future agents should be professionally trained).
- Interpreter: undefined (irrelevant).

2. Aims and Methodology for Experimenting with EMMI

As mentioned from the outset, we wish to test how variations in these main factors — interpreter's speech recognition ability and language proficiency, and all participants' domain knowledge—, will influence the number of disambiguation questions asked by the (automatic) interpreter.

Therefore, in spite of the very small number of experiments we will be able to perform during this stay, we will try to vary systematically some of these main factors.

2.1 Shaping up and pruning the experiment list

Taking into account the previous remarks, let's deploy the main experimental situations in the 'Comprehensive Experiment Configuration Table' below, where only elements < x > or pairs < x , y > appear.

First, Client and Agent ratings language proficiency is reduced to < 6 > in their own language, and to < 1 > in the other language. Second, the domain knowledge of the Agent is reduced to < 1–2 >, that of the Agent to < 5 >, and that of the Client to < 4–5 >.

\ Factor **Subject **	**Language Proficiency** < L_C , L_A >	**Domain Knowledge**	**Auditory Aptitude** < L_C perc., L_A perc.>
Client	< 6 , 1 >	< 4 > < 5 >	< 6 >
Agent	< 1 , 6 >	< 5 >	< 6 >
Interpreter (simulating a machine)	< 2 , 2 > < 2 , 3 > < 3 , 2 > < 3 , 3 >	< 1 > < 2 >	< 2 , 2 > < 2 , 3 > < 3 , 2 > < 3 , 3 >
Total: 4*4*4 = **64 -->**	1*1*4 = **4** situat.	2*1*2 = **4** situat.	1*1*4 = **4** situat.

Remark:

If an Environment Mastery factor were to be taken into account, this would triple, or at least double, the number of experiment patterns.

Let us further reduce the number of experiment patterns.

First, we propose to restrict the Interpreter Domain Knowledge to < 1 >, which corresponds to the most likely situation in realistic IT systems.

Second, the Interpreter's language proficiency can be made symmetrical in case of unequal ability in the two languages used: if the interpreter were to perform a 'unidirectional' translation, we would have to separate between < 2 , 3 > and < 3 , 2 >. But, with the 'bi-directional' process we envisage in these EMMI-based experiments, we propose to consider only one of these two situations, and to denote it by < 2 , 3 >_or_< 3 , 2 >.

This leads to the following 24 of experiment patterns.

\ Factor **Subject **	**Language Proficiency** $<L_C , L_A>$	**Domain Knowledge**	**Auditory Aptitude** $<L_C$ perc., L_A perc.$>$
Client	< 6 , 1 >	< 4 > < 5 >	< 6 >
Agent	< 1 , 6 >	< 5 >	< 6 >
Interpreter (simulating a machine)	< 2 , 2 > < 3 , 3 > < 2 , 3 >_or_< 3 , 2 >	< 1 >	< 2 , 2 > < 2 , 3 > < 3 , 2 > < 3 , 3 >
Total: 3*2*4 = **24** -->	1*1*3 = **3** situat.	2*1*1 = **2** situat.	1*1*4 = **4** situat.

2.2 Human resources and Language pairs

The Domain Knowledge factor will not originate heavy requirements on effective participants to the experiments: it will not be difficult to imagine subject areas meeting the grading requirements.

As far as Linguistic Proficiency is concerned, some regulations should be respected. In order to carry out these bilingual experiments, we ideally need:

- as effective participants, people with a *good* or *very good* proficiency in the language they will speak, and *no* or very *low* ability in the partner's language. This no-ability condition might not be so easy to realize;
- as an 'Wizard of Oz' interpreter, a participant ranking from fluent to expert in both languages. With the bi-directional translation situation, any slight dissymetry in Linguistic Proficiency between the two languages will be acceptable.

Which languages should we consider? In MIDDIM, we have until now focused on Japanese, French and English. After making a quick review of the linguistic abilities of ATR researchers we think might be willing to take part in these experiments, we found that none of them speaks Japanese and no English. Hence, J–E is not considered for these first experiments. For the same kind of reason, F–E is not considered. That leaves us with F–J, for which we are considering as possible participants:

• Japanese	(C or A):	K.H. Loken-Kim, H. Lucke (?)
• French	(C or A):	G. Fafiotte, H. Blanchon
• Interpreter	(I):	M. Tomokiyo, M. Seligman.

2.3 EMMI adjustments

These first experiments will be conducted on the current EMMI environment, without any special extension or modification.

Let us however mention some remarks which have occurred to us after a first trial of the system, and which we hope these experiments can help validate.

1) It seems highly desirable to make available rapidly a sound level control on both Client and Agent stations, in order for them to be able to reduce, or adjust properly, the level of the speech of the other party:

From experience, it seems that in a 3-parties dialogue, a complete blanking of the other end speaker's voice is not at all appropriate. In any case, each dialogue partner should speak only with the

Interpreter. It may even not be helpful — if not outright disturbing, that the Client hears the Agent's voice at the full level, and vice versa.

The Interpreter might even be given some access right to this control adjustment, in order to better manage talkative confusing situations.

2) We also lay great stress on the importance of different types of feedback to be given to the end speakers (Client, Agent) in a multimodal manner.

We will not here anticipate the results of Dialogue-Based Disambiguation experiments regarding best-suited user feedback. We think that a methodical experimentation of the potential role of various kind of feed-back, at the two ends, should be taken as part of a prototyping paradigm.

For instance, how to help in regulating the dialogue and in promoting some spontaneous well-disciplined listening attitude? Some thin colored line could appear on the screens, or an icon could flash, etc., as soon as someone (A, C or I) is talking, or depending on the Interpreter's preferential attention to one or the other party. The (machine) Interpreter would in some cases operate a switch to express this 'turning towards' attitude or 'oriented expectation'.

3. Tentative experiment planning

3.1 Experiment 1 (4 phases): explaining how to program the Viterbi algorithm

\ Factor **Subject \**	**Language Proficiency** < F , J >	**Domain Knowledge**	**Auditory Aptitude** <F perc., J perc.>
H. Blanchon C	< 6 , 1 >	< 4 >	< 6 >
K.H.Loken-Kim A	< 1 , 6 >	< 5 >	< 6 >
M. Tomokiyo I (simulating a machine)	< 5 , 6 > to be compensated down to < 2 , 3 > by vocoder	< 1 >	< 1–2 , 1–2 > instead of < 2–3 , 2–3 > through vocoder distortion

The situation should be that A explains to C how to program an algorithm. Here, A should explain to C his program for the Viterbi algorithm.

We envisage 4 successive sessions, with a different aspect being explained each time, so that the participants cannot learn and improve their knowledge of the subject matter from one phase to the next. In each experiment, we will try to simulate decreasing levels of the (machine) interpreter's auditory aptitude by increasing the distortion of the vocoder.

Phase	**Topic**	**Perception levels (I)** < F , J >	**Remarks**
1	Explanation of the algorithm (goals, method)	< 2 , 2 >	Distortion set to compensate for too high linguistic ability
2	Explanation of the data structures used	< 2 , 1 >	Japanese side more distorted
3	Explanation of the coding of the algorithm	< 1 , 2 >	French side more distorted
4	Explanation of a run of the program	< 1 , 1 >	Both sides more distorted

It should be possible that A shows the text of the program and of the trace of a run on the screen, so that A and C can point at it.

Each session should last 20 minutes or less, as interpreting is a rather tiring task. Thus, it should be possible to record the 4 phases on one 90 mn video tape.

3.2 Experiment 2 (3 sessions): explaining how to go to the Grenoble Culture House from Lyon-Satolas airport

\ Factor Subject \	Language Proficiency <F , J >	Domain Knowledge	Auditory Aptitude <F perc., J perc.>
T. Morimoto C	< 1 , 6 >	< 3 >	< 6 >
G. Fafiotte A	< 6 , 1 >	< 5 >	< 6 >
M. Seligman I (simulating a machine)	< 4 , 4 > to be compensated down to < 2 , 3 > by vocoder	< 2 >	< 1–2 , 1–2 > instead of < 2–3 , 2–3 > through vocoder distortion

The situation should be that A explains to C how to go somewhere. Here, A should explain to C how to reach the Grenoble Culture House from Lyon-Satolas airport.

Again, we envisage 3 successive sessions, with a different aspect being explained each time, so that the interpreter cannot learn and improve his knowledge of the subject matter from one phase to the next. In each experiment, we decrease the auditory aptitude of the interpreter by increasing the distortion of the vocoder.

Phase	Topic	Perception levels (I) < F , J >	Remarks
1	Explanation of how to come to Grenoble bus station	< 3 , 2 >	Distortion set to compensate for too high linguistic ability
2	Explanation of how to go to the tramway station	< 2 , 1 >	Japanese side more distorted
3	Explanation of how to go to the Culture House	< 1 , 1 >	Both sides more distorted

3.3 Exploitation of the results

The resulting two video tapes will be transformed into QuickTime files on a Macintosh. In principle, 7 files should be produced, corresponding to the 7 sessions.

The number and duration of the clarification (speech) or disambiguation (language) questions asked by the interpreter will be measured and compared with the number and duration of all utterances.

If possible, the study will be refined according to the ambiguity types proposed in M. Tomokiyo's MIDDIM report.

If time permits, we would also like to prepare one more batch of experiments, to be conducted during this stay, this time tackling the J-E language pair. This would however necessitate the participation of one or two Japanese persons having no working knowledge of English.

Acknowledgments

We are very grateful to all the above-mentioned persons for their helpfulness and readiness to spend some time on these experiments, as well as to Dr. Y. Yamazaki and T. Morimoto, for their constant support to this project.

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