A SIMULATION AND COLLECTION PLATFORM ON THE NET FOR MULTIMODAL TRANSLATED SPOKEN DIALOGUES

Georges FAFIOTTE

CLIPS-IMAG / GETA

UJF-Grenoble 1 University, CNRS – BP 53

38041 - GRENOBLE Cedex 09 (France)

georges.fafiotte@imag.fr

KEYWORDS

Speech Machine Translation, speech corpora acquisition, spontaneous multilingual spoken dialogues, multimodal speech processing, network-based platform, SMT simulation, Wizard of Oz.

ABSTRACT

We developed the SIM* environment, dedicated to various applications in multilingual multimodal spoken dialogue processing, and a companion project to the international C-STAR and European NESPOLE! programmes.
The first SIM*/1 system is multiplatform (Mac OS, Windows, Unix), and operates under a Wizard of Oz scheme, as a simulator for supporting and collecting bilingual spontaneous spoken dialogues between distant partners through local networks, while simulating adjustable quality interpreting. It was first tested as a collecting environment to build raw speech corpora in French-English on small dialogues close to C-STAR scenarios. It is currently used for collecting bilingual speech corpora in realistic spontaneous task-oriented translated French-Chinese dialogues (travel information and reservation, hotel reservation for business trips).
Under current development now is SIM*/2, an Internet-supported version with functional extension towards multimodal interaction. Some multimodal features are here available, such as interactive marking on shared local whiteboards, proper nouns or keywords spelling, user-driven file transfer, on line extraction of Web information, simple video-conferencing.
Another aim for the platform is speaker observation, with on-the-fly capture of linguistic, multimodal, or behavioural events. Besides, SIM*/2 is intended to run as a plug-in platform for testing and tuning different components in SMT, with either the presence of human Wizard interpreters or automated processing.

INTRODUCTION

Building speech machine translation systems for multimodal dialogues is very important for easing and promoting international domain-oriented oral exchange on the web, particularly for e-business and teleservicing. We therefore need large multimodal multilingual speech corpora, with enhanced transcriptions.
We are here reporting work that stands somewhat upstream of the actual modelling and testing of annotation multimodal/multimedia schemas, and accounts for the building under way of a platform contributing to an operative framework on the field.

Year by year Speech Machine Translation (SMT) applied to spoken realistic dialogues on restricted domains is showing rapid improvement, in particular towards style spontaneity, multilingualism, due to dedicated effort such as the C-STAR project (international Consortium for Speech Translation Advanced Research) [Boitet & al., 98].
However we are still facing a strong need for collecting actual spontaneous multilingual spoken dialogues, for building large annotated speech corpora, and producing specialized or general lexical bases, in order to efficiently train SMT components.
Parallel to our integration commitment of the French language within C-STAR II, we aimed at developing a collection and experimentation platform to meet such requirements and somewhat broaden them. In the context of our research on spoken dialogue processing, we are inspired by both previous experiments on the multimodal Wizard of Oz EMMI platform at ATR-ITL [Loken-Kim & al., 94] and by the multi-wizard NEIMO platform at CLIPS [Coutaz & al., 93].
The SIM* project takes interest in prototyping a multipurpose adaptive simulation environment for spoken dialogue processing. The system is to provide such facilities as multilingual speech corpus collection, simulation of 'real life speech' translation, observation and capture of speakers spontaneous attitude in multimodal settings. We aim at developing both LAN- and Internet-based speech collecting platforms, and at experimenting different architectures, in order to enhance genericity equally towards multimodality, multiplatform implementation, differential lingware resource plug-in, multilingualism.

We first present the research context and motivation for the simulator, before sketching a generic architecture. Following, the paper presents the SIM* environment, its first use for bilingual spoken dialogues collection, and development under way. We then draft perspectives and issue concluding remarks.

MOTIVATION AND GENERIC ARCHITECTURE

Research and project context

In earlier work at ATR-ITL in Nara (Japan), in the context of Multimodal Interactive Disambiguation (MIDDIM common project) [Loken-Kim & al., 96], we designed and ran a few bilingual speech translation pilot-experiments (Japanese-English, Japanese-French) on the EMMI multimodal Wizard of Oz platform [Fafiotte & Boitet, 96]. As known [Salber & Coutaz, 93], in a Wizard of Oz environment a hidden human is answering users requests in place of functional software components (here an bilingual interpreter), in order to simulate, to remedy or to enlarge computer resources, to collect data, and to possibly observe and capture user reactions during experimentation.

We then spontaneously derived a manual annotation scheme, similar in spirit to multistave musical scores, to tick-off down on paper main fly-on-the-wall traits. We used it to notate both multimodal events (available then) and linguistic events typical to oral expression (such as clarification subdialogs, false starts, hesitations…), in the course of dialogues.

We also inherit the spirit of generic aspects in NEIMO — a multi-wizard of Oz environment which was intended to observe and analyze monolingual multimodal interaction, with Telecoms Multi-servicing as a host-application.

Last our motivation actually is close to the effort of the C-STAR and NESPOLE! communities, which both aim at demonstrating multilingual quality Speech Machine Translation through clustered demos on service-oriented subdomains. C-STAR II currently handles six languages (English, French, German, Italian, Japanese and Korean), with an interlingual scheme using a task-oriented common Interchange Format representation. Within the framework of these programmes, numerous multilingual spoken dialogues are to be collected, in order to train basic recognition and translation functions for the speech translation demonstration platform.
We actually think of SIM* as a complementary approach: a multipurpose network-based multimodal multilingual simulation platform.

Objectives

A versatile simulation platform:
Goals were assigned to SIM*, to be stepwise achieved:
– to simulate speech interpreting of real spontaneous dialogues, possibly with adjustable quality interpreting (from high-performance to still non-perfect linguistic and/or acoustic quality …of the Wizard translation), in order to experiment later on speaker specific subdialogic adaptation or rejection,
– to collect multilingual speech corpora from dialogues between monolingual native speakers,
– to add multimodal features to C-STAR basic interaction and to experiment on it in various situations,
– to allow the testing of various lingware components through differential plug-in of software and/or Wizard resources,
– to enable observation and capture of speakers activity, of multimodal and speech-relevant linguistic events, of behavioural factors.
– to develop automated aids to high quality annotation of multimodal speech corpora.

A network-based environment, first to be experienced and assessed on an Intranet:
The platform should support real-life bilingual dialogues, possibly between far distant speakers. Hence communication and control engines should be Internet-supported, though being first implemented on local networks. In order to improve the exchange rate during collecting sessions, speech signals can be, similarly to the early EMMI platform, locally sent through wired audio lines, for instance between close rooms in the laboratory. Speech files transfer though should be network-driven indeed, for distant conversing.

Linguistic processes to be carried out:
The first main use for the SIM*/1 system is the gathering of large task-oriented bilingual corpora of spontaneous speech, between very or not very distant monolingual native speakers. We are currently starting to collect Chinese-French dialogues.
We are well motivated by an assistance to corpus annotation, and how to integrate the enrichment of primary transcriptions or raw automatic annotations. Another process related to data acquisition is the capture on-the-fly of multimodal events, to be resynchronized with translated utterances.

Open generic architecture

Multilingualism on the simulator is to be expandable: language genericity is actually ensured by the architecture, around one or several Wizard of Oz interpreters (cf Fig. 1).

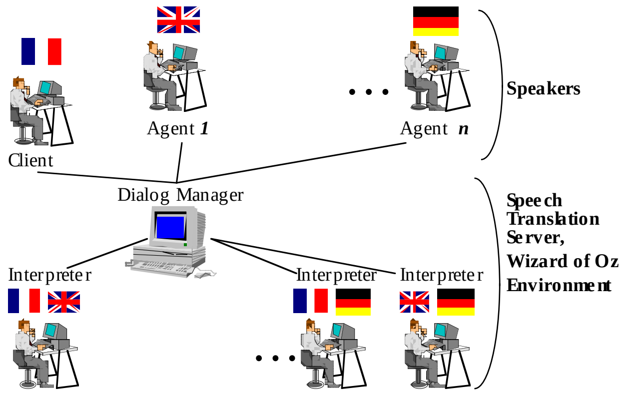

Fig. 1: SIM architecture in a multilingual version*

We wish functional genericity as well, namely to provide simple reconfiguration of main functions (recognition, translation, synthesis): for each such function, to be able on request to choose either software processing or human-driven (Wizard of Oz), or else to test alternative lingware functional components.
Physically the environment is to be an open multiplatform, running as well on heterogeneous station configurations (Mac, PC, Unix).

SIM*/1 FUNCTIONALITIES, TECHNICAL ASPECTS, PRESENT USE AND DEVELOPMENT

Main functionalities

In the first LAN-based SIM*/1 implementation, two monolingual speakers converse in possibly distant rooms in the laboratory. The current platform carries out (cf Fig. 2, 3):

– Wizard of Oz interpreting technology, to simulate Speech Machine Translation of bilingual dialogues,
– a communication and control engine based on a client-server scheme, in charge for speech turns sequencing and dialogue regulation,
– in the collect-oriented mode, the building of bilingual session bases (regarding each speaker, the interpreter), for speech and textual events with dating and identification attributes for every utterance,
– some available multimodal features, such as interactive marking on shared local whiteboards, proper nouns spelling and text editing, some video-conferencing, user-driven file transfer, on line extraction of Web information to illustrate or stimulate dialogues.

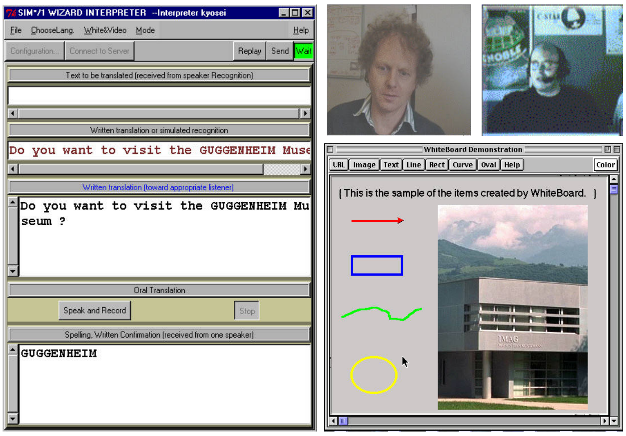

Fig. 2: Wizard Interpreter station

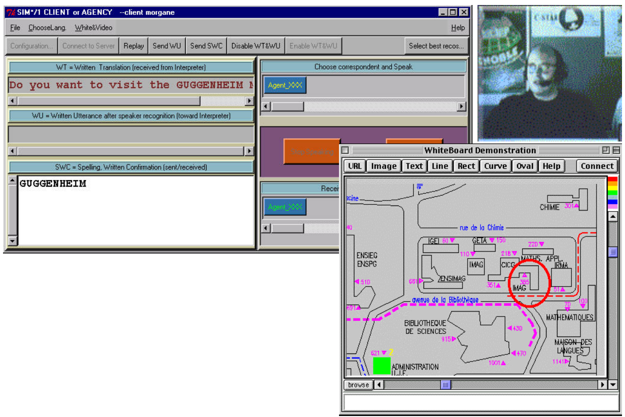

Fig. 3: Client (or Agent) station

A typical dialogue situation currently demonstrated is as follows (cf Fig. 4):
– two (or more) monolingual speakers converse in video-conference (Client and Agent, in English and French),
– speech events are recorded on the appropriate speech bases then transmitted to the Wizard Interpreter,
– an oral or written translation (depending on initial setting) is issued by the Interpreter, recorded on a session base, then sent to the addressee,
– if needed vocal synthesis of the Wizard translation is done on addressee's workstation.

Some automatic speech recognition will be integrated, first to the Interpreter workstation, with display and eventual confirmation among recognized utterances, or spoken correction. The speech of a trained Wizard Interpreter is expected to sound clear and clean.

Speech recognition for the speakers will then take place, once sufficient domain-specific data may be bootstrapped in the process.

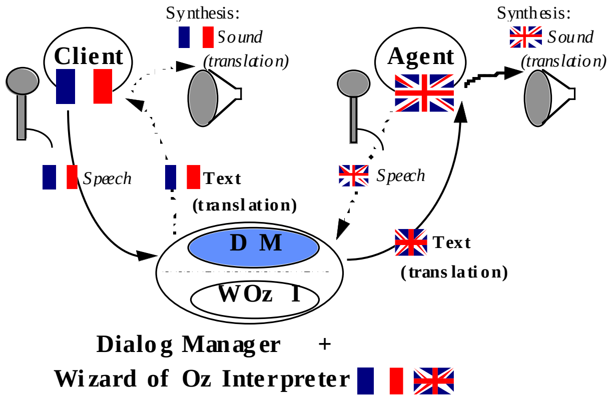

Fig. 4: Principle of the prototype first version

On the current implementation both the Dialog Manager and the Wizard Interpreter processes may run on one same station.
Recent development has explored collecting-oriented variants (for monolingual dialogues, or with bilingual speakers while by-passing any translation process), or 3-speaker situations (1 Client, 2 Agents), anticipating the study of specific aspects in the translation of multiple dialogues.

Current use in collecting speech corpora

Early use while testing SIM*/1 includes monolingual dialogue collection, in the context of the C-STAR II project. Collecting mode is currently used in gathering bilingual spontaneous spoken dialogues between monolingual Chinese and French speakers, on task-oriented domains (travel and hotel information and reservation for business trips), in cooperation with the Natural Language Processing Research group (Academy of Sciences of China, Beijing).

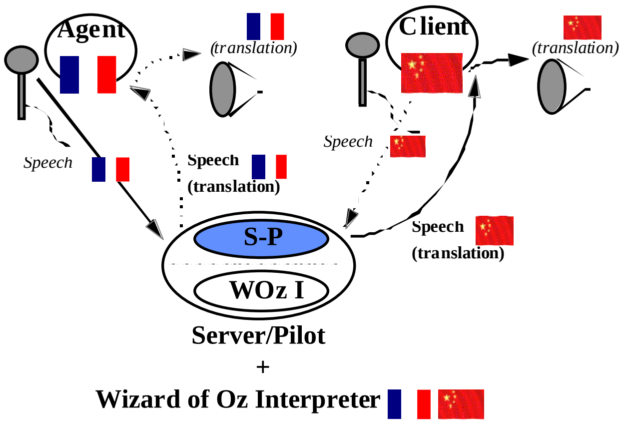

Fig. 5: Current version of the SIM prototype, as being tested for collecting Chinese-French spontaneous dialogues on travel/hotel reservation (multiplatform implementation with, for both Server/Pilot and Wizard of Oz Interpreter, either one same Unix platform or two separate PC and/or Macintosh stations)*

Technical view

SIM*/1 work stations indifferently run on any platform from Apple Macintosh, PC or Unix stations.
Development tools include both TCL-TK multiplatform and MBone applications, which implement client-server servicing, data recording, speakers and interpreter interfaces, including the video-conferencing facility.

Towards a transcription and annotation assistant

Lately we initiated development plans to model additional aids, based on generic multilevel dependency parsers, to produce 'quality' transcripts or enhanced annotated corpora, namely to correct primary automatic transcripts of multimodal multilingual dialogues, and to complement them somehow at semantic and pragmatic levels.

EVOLUTION

The SIM*/1 primary platform gradually meets main early requirements. First use gives a fair account for its potential versatility. It still needs be improved according to different scopes of interest: technical assessment for feasibility, performance on Internet-based real-life distant dialogues, on data acquisition, on converser observation with different hypotheses on multimodal interface.

Development towards SIM*/2 will progressively include:

- Long distance use: coming development will lead to remote connections over the Internet in order to work out better realistic situations.
  Transition to net-supported dialogues easily derives from the current implementation. Up to now, both drawbacks of distant live speech transmission on the Internet (with still uncertain reliability when using regular connection), and the prohibitive cost of ISDN connections, may require for some time telephone parallel dialogue transmission, with locally recorded speech segment files being transmitted in a row, at a slower pace.
- Enhancing plug-in facilities, and adding an efficient start-up configuration resource.
- Experimenting on clustered sub-domains, partially with MT automatic and with Wizard processing.
- A more generic communication engine, compatible with related platforms (NESPOLE!, C-STAR III).
- Extending multimodality: among other points, provide for resynchronization between oral translation and multimodal events (pointing, marking, spelling…).
- Integrating specialized 'Wizard observer': for capturing multimodal events, prosodic or speech-relevant linguistic events (in order also to study particular features relevant to multilingual dialogues, anaphorae in multispeaker situations…).
- Producing on-line on-the-fly transcriptions and annotations (eventually involving extra dedicated Wizard stations), towards automated ones.
- Expanding multilingualism: due to the architecture, generalization to the simulation of multilingual interpreting should follow quite easily, beyond logistic and session protocol aspects to be settled.
- Studying both interpreter or speaker, instant or asynchronous, multilingual lexical aids.

CONCLUSION

As undergoing development in multilingual Speech MT technology sparks off a need for methodically building large specialized or general annotated speech corpora, Internet appears to be a stimulating and efficient vehicle for collecting spontaneous multilingual oral dialogues.

We presented here first achievements and work in progress on SIM*, a multipurpose simulation and collection environment, under experimental use for data acquisition.
But since we really stand at the beginning of the process of collecting and annotating multimodal bilingual speech corpora, we are willing to experience various approaches in meta-descriptions and annotation schemas, and to integrate and try out available models and techniques.

There is a growing interest as well in observing and modelling speakers behaviour or expectation, in the situation of multimodal multilingual machine-aided dialogues over large networks, whereas multimodal portable communication devices currently spring up. Wizard of Oz technology proves to be efficient for such observatory investigation, in order to enhance human-aided MT.

We also expect such an open simulation platform to contribute to explore new types of Internet- or LAN-based multilingual applications. As a illustration case we precisely think of a new involvement for professional or occasional human interpreters intervening via the net.
Parallel development is actually in progress towards a SIM*/3 variant platform, involving a symmetrical paradigm: the system behaves as an assistant to a real human interpreter, operating on demand. The interpreter would remotely assist interlocutors (two distant speakers, or a video-conference group), who otherwise manage to converse in a common language, and ask for assistance during sensitive parts of the conversation. The system will provide the interpreter with instant automatic lexical or terminological backing or follow-up, derived from SMT active functions monitoring the preceding dialog phase (on a specialized domain). Of course such an intermittent interpreter could benefit asynchronous aids of this type only while strongly regulating them, in order to avoid inappropriate perceptive or cognitive load.

ACKNOWLEDGEMENTS

Reported work is backed by UJF (Université Grenoble 1), INPG, CNRS, and granted in part by the Rhone-Alpes Region (ERIM convention, Emergence 99' programme), and by a LIAMA convention (Beijing- Paris).
The author is most grateful to ZHAI JianShe (Visiting Researcher from Nanking University) who patiently developed successive prototypes, and to Laurent AUBLET-CUVELIER (from the MULTICOM team at CLIPS) for his contribution to early specification and experimentation tuning. Thanks also to many members of the GETA and GEOD teams at CLIPS-IMAG, for spontaneous acting as experimentees on the SIM* simulator.

REFERENCES

http://www.c-star.org   C-STAR official site.

Boitet C. (1997) *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects.* Proc. PACLING-97, Ohme, 2-5 Sep. 1997, Meisei University, pp. 23-57 (invited communication).

Boitet C., Caelen J., Fafiotte G., Keller E., Lafourcade M. & Wehrli É. (1998) *Integrating French within C-STAR II: second report & demos of the CLIPS++ group.* Proc. C-STAR international meeting on Speech Translation, Grenoble, 12-14 Jan. 1998, CLIPS-IMAG, 16 p.

Coutaz J., Salber D., Carraux E. & Portolan N. (1996) *NEIMO, a Multiworkstation Usability Lab for Observing and Analyzing Multimodal Interaction.* Proc. CHI'96 Companion, 1996, 2 p.

Fafiotte G. & Boitet C. (1994) *Report on first EMMI Experiments for the MIDDIM project in the context of Interpreting Telecommunications.* MIDDIM report, TR-IT-0074, GETA-IMAG & ATR-ITL, Aug. 1994, 11 p.

Fafiotte G. & Boitet C. (1996) *An Analysis of the first EMMI-based Experiments on Interactive Disambiguation in the Context of Automated Interpreting Telecommunications.* MIDDIM-96 Seminar, Le Col de Porte, Aug. 1996, pp. 224-237.

Fafiotte G. & Zhai J.S. (1999) *A Network-based Simulator for Speech Translation.* Proc. NLPRS 99', Beijing, 5-7 Nov. 1999, pp. 511-514.

Loken-Kim K.-h., Boitet C. & Morimoto T. (1996) *A chronicle of ATR-GETA multimodal interactive disambiguation research collaboration.* Proc. MIDDIM-96 Seminar, Le Col de Porte, 12-14 Aug. 1996, pp. 23-28.

Loken-Kim K.-h., Yato F. & Morimoto T. (1994) *A Simulation Environment for Multimodal Interpreting Telecommunications.* Proc. IPSJ-AV workshop, Mar. 1994, 5 p.

Salber D. & Coutaz J. (1993) *Applying the Wizard of Oz Technique to the Study of Multimodal Systems.* Proc. of East-West HCI 93', Aug. 1993, 12 p.