Robust Spoken Translation at ITC-IRST

Gianni Lazzari
ITC-IRST
Trento, Italy

Abstract

In this paper the ITC-irst research issues and approach to the spoken translation problem will be presented together with a description of the demonstration system developed in the framework of C-STAR II Consortium. The challenge of future applications in the e-commerce and e-service sectors will also be presented and discussed.

1 Introduction

Significant progress has been made in the field of human language technologies. Various tasks like continuous speech recognition for large vocabulary, speaker and language identification, spoken information inquiry in restricted domains are today feasible and different prototypes and systems are running. The spoken translation problem on the other hand is still a significant challenge[16]. Nevertheless recently important advances have been demonstrated approaching this problem as a human to human communication task carried on in a restricted domain. In this case bi-directional, real time operation is necessary, but fairly low quality is acceptable when communication is achieved. A Spoken Translation System does not need to give a complete correct solution, if it produces a sufficient expression in the target language satisfying the human dialog situation. If the human to human communication is also supported by multimedia information this approach is even more convincing. This solution to the problem of spoken translation is also suggested by the needs of the information society. The diffusion of Internet and the related services is asking for high sophisticated tools for human to human communication over the Net. Web Phone. Video Call Centers, Call Centers, Videoconferencing Systems allow today the possibility of a global communication infrastructure, despite the lack of adequate band. This infrastructure offers high human to human communication capabilities with the only exception of the language, which is and will be for the future a significant barrier to overcome.

ITC-irst has been working on spoken language technologies since 1987 and on spoken translation since 1995, when joined the C-STAR II consortium as a partner. The approach chosen and pursued at ITC-irst to the spoken translation problem is based on the following assumptions:

- optimize the capability to transmit the meaning of the communication focusing on the robustness of the system intended as capability to manage spontaneous speech and ill formed language (the so called ungrammatical spoken language) and on aid of multimedia information.
- focus on the scalability and portability of the system. This means to be able to extend a given domain and/or to port the system to different domains with reasonable effort.
- demonstrate the usefulness of the use of 'interlingua' or 'interchange format' approaches.

Considering the translation task as a coupling of two different passes, an analysis step which produces a suitable interlingua from a given speech signal, a generation step, which produces a spoken sentence in the target language starting from an interlingua representation.
In this paper first of all the ITC-irst research issues and approach to the spoken translation problem will be presented together with a description of the demonstration system. Then the challenge of future applications in the e-commerce and e-service sectors will be presented and discussed.

2. Research Issues

The scientific and technological research issues we intend to address in order to improve over current experimental **speech-to-speech translation** (STST) systems, are: **robustness, scalability, cross-domain portability** and **multimodal interaction** with **multimedia content.**

2.1 Robustness

In order to fully support natural interaction, the system must be able to cope with the disfluencies of spontaneous speech including interruptions, corrections, repetitions, false starts, etc. In these respects, an essential feature of the system will be its robustness — that is, the capability of dealing with corrupted inputs, due to either the peculiarities of the input utterance or to errors of the acoustic recognizer, and with incomplete information. Robustness is so important to STST systems that it is usual practice to sacrifice fidelity to the input utterance in favor of a smooth continuation of the communication process. The system is required to be able to properly translate at least all the input content, which is relevant to the accomplishment of the domain task(s), this way simplifying the goal with respect to a complete and precise translation of the input utterance.

Robustness has to be assessed, both for the complete system and for the single speech/language modules, by using real speech data.

2.2 Multimodality

In order to be really effective for negotiation in e-commerce/service, the system cannot simply handle speech, language or text. E-commerce already heavily relies on images, graphics and animations describing the products. A multimodal support has to be provided for negotiation by allowing a close integration of, and interaction between speech-based communication and visual cues and content. Both the customer and the provider will be enabled to talk about content that is presented by means of images and make reference to it, while speech translation will resolve the reference that the two partners make to visually presented material. At the same time, the system will provide supporting information dynamically, as the human-to-human dialogue unfolds. Many of these goals are still research topics in human-machine communication scenarios. They are particularly challenging in a translingual human-to-human communication setting, where multimodal coordination and synchronization must be maintained during and across language translation.

2.3 Scalability and cross-domain portability

Present STST systems are highly domain dependent. In order to improve effectiveness and performance, both the linguistic resources and many of the speech/language engines are built around a particular domain at hand. This results in systems, which do not easily scale up, and are difficult to port from one domain to another without major changes. Although attempts at addressing **scalability** and **cross-domain portability** have already been made in many areas of HLT, these are rather new concerns in STST. They are crucial, though, to offer STST as a viable technological solution. Such problems will be addressed by improving the performances of single engines architectures: enhancing existing Intermediate Representation Formalism (IRF) to provide for more flexible, easy to use and to extend domain and meaning encoding.

2.4 The Interchange Format

The adoption of an Interchange Format (IF) based approach has several advantages and potentialities. The most obvious advantage is the reduction of the number of different systems, which have to be implemented. Given *n* different languages, an analysis chain (starting from the spoken input and delivering an IF representation) and a synthesis chain (taking the IF representation and providing a linguistic form for it) for each language suffice to yield a system capable of dealing with speech-to-speech translation between all of the possible language pairs. That is. the resulting system would require *n* separate analysis and synthesis chains, instead of the otherwise required quadratic number of modules. Furthermore, given that each module involves only one language, native speakers of that language can do the development. Another important advantage concerns portability to a new language; given the described configuration, a lower effort is necessary to make an existing system capable of dealing with a new language. This strikingly contrasts with the case of a direct translation system, where the addition of a new language to a set of *n* pre-existing languages requires the construction of n new complete modules to link each old language to the new one. In addition to this, the techniques developed to build and process a formal representation of the information content of utterances can be exploited to meet the demands of many other applicative scenarios. For example, in a speech based information retrieval system, the IF can be used to build the formal query. Both in this case and in speech-to-speech scenario, the IF representations can provide the means to produce summaries of the transactions occurred between the human and the machine, and among the different parties, respectively.

For the IF based approach to work properly, the IF design is crucial. This is a difficult problem because many aspects need be taken into account. In the first place, the IF must be (as) language independent (as possible). That is, we want it to focus on the information contained in utterances rather than the way in which such information is expressed. On the other hand, robustness requires that the IF be able to also code partial information. Another important point is the dependence of the interchange format on domain: this is obviously a limitation, but the applicative scenario provides strong indications as to the type of information which must be extracted, while the domain defines the actual data to be identified in the utterance. For example, the application might request the identi-

fication of the topic of the utterance, and the domain might restrict the choice to a finite set [11]. Speech-to-speech translation scenarios feature dialogues with a richer structure than that of speech-based information retrieval and the interchange format must be able to capture such structure. For instance, the number of parties active in the conversation is always greater than one and track must be kept of who is speaking at each time. To this end, the adopted IF provides a label to encode the speaker: a: for the agent and c: for the customer. Furthermore, the users do not only ask for data or information, but they also perform, and request the other party to perform, a number of different actions. Users greet the other speaker, seek and give information and clarifications, accept and reject proposals and suggestions, and so on. All these actions are represented in the speech-act level of the adopted IF. Moreover, each action may concern and involve a number of objects in the world and properties thereof. Such objects and properties are encoded by means of concepts, usually ordered by importance and (decreasing) generality. Examples of the relevant concepts are availability, price, hotel, room, trip, sight-type, ...
Finally, there is an argument level, consisting of attribute-value pairs such as time=sunday, location=downtown etc. Each concept admits one or more attribute-value pairs. In summary, the IF consists of four levels:

- the speaker label
- the speech-act part
- a (sequence of) concept(s)
- the arguments.

The architecture of the IF permits to clearly distinguish the domain dependent part (concepts and arguments) from the domain independent one (speaker and speech-act). This facilitates the porting from a domain to another. For example, moving from the hotel reservation domain to the travel information one will only require the addition of new concepts and new arguments.

3 System Architecture

The ITC-irst system architecture consists of two main processing chains: the analysis chain and the synthesis chain (cf. Figure 1). The analysis chain converts the Italian input speech signal into a (sequence of) IF representation(s) by going through: the *recognizer,* which produces a sequence of word hypotheses for the input signal; and the *understanding module,* which exploits a multi-layer argument extractor and a statistical based classifier to deliver IF representations. The synthesis chain starts from an IF expression and produces an Italian synthesized audio natural language message expressing that content. It consists of two modules. The *generator* first converts the IF representation into a more language oriented representation and then integrates it with domain knowledge to produce sentences in Italian [12]. Such sentences feed a speech *synthesizer,* namely the text-to-speech system Eloquens developed by CSELT.
All translation systems developed by C-STAR partners have to communicate to each other. A commumcation server performs the interface between the ITC-irst system and the other ones.

3.1 The Acoustic Recognizer

The goal of the speech recognizer is to find the word sequence that most likely caused the acoustic signal. For each sentence hypothesis, the likelihood can be computed on the basis of two probabilistic models: the Acoustic Model (AM), concerning the relation between each word and the acoustic signal, and the Language Model (LM), which considers how words are concatenated in a sentence. Speech sound is first converted into electrical signals by a microphone; the analog signal is then digitalized through sampling and quantizing: the digital signal is finally processed to estimate significative parameters [10]. The recognizer used in our system is based on the platform developed at ITC-irst [7] for Italian large vocabulary dictation tasks. It uses an AM based on hidden Markov models (HMMs) [13] of phonemes. Each vocabulary word is modeled by concatenating the sequence of phoneme HMMs corresponding to its pronunciation. HMMs were initialized on a phonetically rich database, APASCI, collected at ITC-irst [1] while a re-training phase was carried out on the training set of the spontaneous speech corpus collected at ITC-irst for the C-Star project. Specific models were trained for most frequent extra linguistic phenomena. The LM is based on trigrams [7] and includes extra linguistic models. AM and LM are compiled into a finite state network that defines the search space, a: phoneme level, for the Viterbi-based decoding algorithm [3].

3.2 The Understanding Module

The input to the understanding module is represented by a sequence of words and speech phenomena like pauses and hesitations corresponding to a semantic unit, that is. a part of the utterance identified by two semantic boundaries. The role of the understanding module is to construct the corresponding IF representation. The first level of the IF representation, i.e. the speaker, is known. The rest of the IF representation is built in two successive phases, each devoted to a single task. In the first phase the utterance is processed by a multi-layer argument extractor [14] based on Lex and Yacc. In a second phase, the speech act and the concept sequence is constructed.

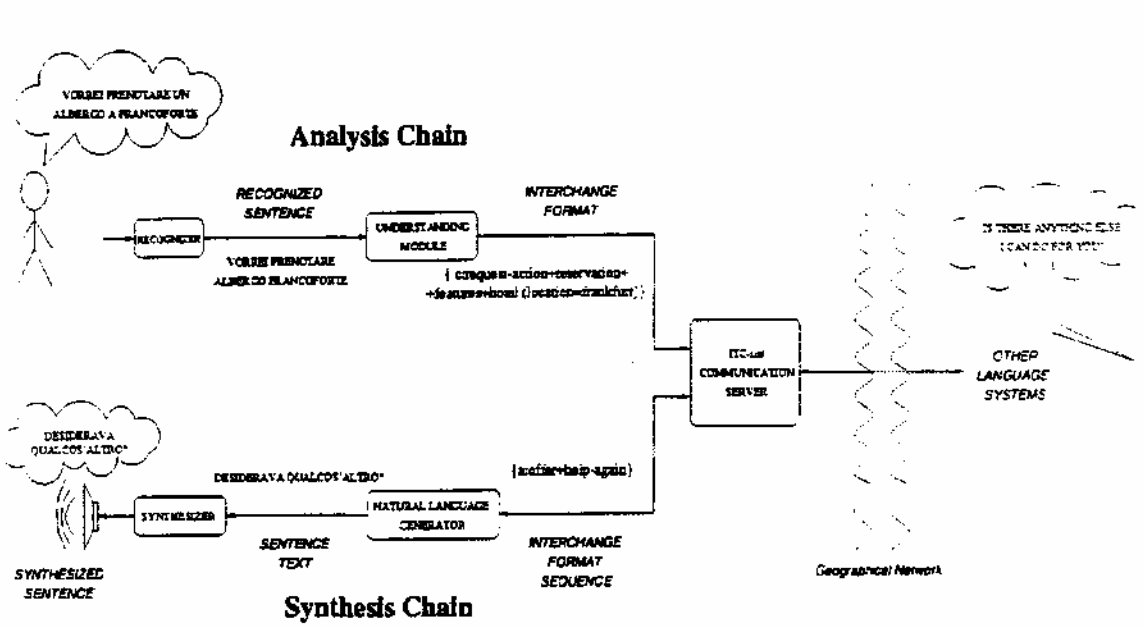

Fig. 1 – ITC-irst system architecture

Its structure is described by a regular grammar, represented by an a-cyclic graph. Scores are computed for each arc in the in the graph by using Semantic Classification Trees (SCTs) [6], which are derived from CARTs (Classification and Regression Trees) [2]. SCTs can be automatically built from a corpus of labelled examples (training set). In the current implementation [5], a variant of the procedure described in [6], the classification is based on the presence of particular keywords. In a way, keywords are the simplest and most straightforward elements, which can be identified in a sentence, while in principle statistical classifiers could be applied to any structure built on the input. On the other hand, in the system here considered, the application of SCTs is preceded by the application of the multi-layer argument extractor which, in addition to extracting the arguments, preprocesses the input data by substituting important phrases with labels which can be used as keywords. In this way, the statistical relevance of the data improves because equivalent phenomena are clustered in the same event.

3.3 The generator

The goal of the generator is to produce the Italian translation for the source sentences encoded in IF representations. Since IF is in not a linguistically oriented semantic representation, first the generator uses a sentence planning module mapping IF into a functional representation similar to the f-structure of Lexical Functional Grammar. This step usually does not take place in current machine translation systems.

The sentence planning algorithm allows for three strategies, which can be mixed. General heuristics are used to map parts of the IF representation in linguistically relevant functions. These rules give medium quality results, but can be applied in a large number of cases. However, before resorting to these general rules, the system can consider more specific ones, which produce a more accurate output but can be applied in a smaller number of cases. Should both strategies fail, a backup strategy applies, which merely translates the components of the IF representation, see [12]. The actual generation step makes wide use of flexible templates, which guarantee high efficiency and at the same time allow us an elegant treatment of linguistic phenomena such as phonological adjustment, morphological agreement and syntactic constituency.

3.4 ITC-irst system demonstrator

The ITC-irst speech translation system has been demonstrated a number of times during 1999. The demonstration concerns an Italian traveler who is connected simultaneously to the US travel agency in New York and to the German travel agency in Frankfurt. The three people can communicate by speaking their own language thanks to the translation systems, while they can see and hear each other through a 3-point video conference. During the demonstration, the ITC-irst system is located in a conference room where both the video conference images and some information shown in Figure 2 are projected on a large screen. In this way, the audience can follow the dialogues not only hearing them, but also seeing the texts in Italian.

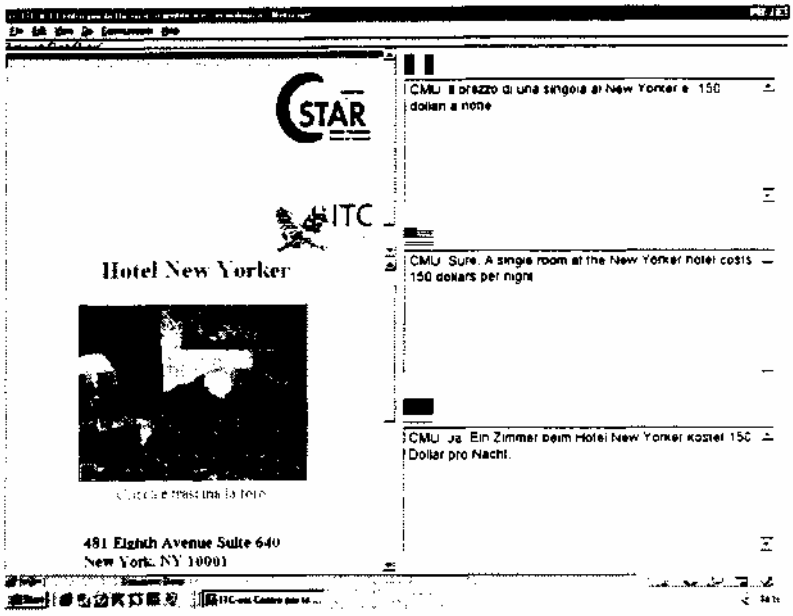

Fig. 2 - ITC-irst demonstrator associates speech-to-speech translation with WEB images

English and German, and the Web pages sent by agents to the traveler. The semantic of the dialogues regards booking flights and hotels, and the possibility of asking for some simple tourist information, like opening hours and ticket costs of museums: agents can also enrich the information given by voice messages to the traveler, by sending proper Web pages.

4 E-commerce, e-service: a challenge

E-commerce opens the possibility of selling goods and services on a global market supported by the WEB. Up to now, however, interactions are mainly menu based, guiding the potential client through a choice tree allowing only for a limited set of predefined results. This reflects a conception of e-commerce/service as basically driven by the supplier side. As a consequence, identification of customer's requirement and the consequent processes of extending existing product range and product innovation are crucially not supported; on the contrary, the interaction aims at fitting the customer (notice, not the customer's needs) into existing product ranges and configuration options provided by the supplier.
A different perspective on e-commerce/service is the one in which e-commerce/service focuses on the identification of customer's needs, being prepared to dynamically configure and propose complex solutions. In such a scenario, the customer and the provider are involved in a **negotiation** - the process where the client is allowed to explore alternative solutions, while the e-service/commerce provider reacts by eliciting unexpressed needs and configuring viable alternatives. Human-to-human interaction can obviously greatly contribute to achieve such objectives. Humans, in fact, can fully and naturally understand other humans' motivations and desires, can answer their needs in creative ways, possibly without fully matching the customer's requests, but always trying to find solutions that maximize the customer's satisfaction vis a vis concrete possibilities. In a word, humans are naturally equipped for the negotiation task.
The goal of making e-commerce/service capable of supporting negotiation requires that we provide for systems capable of dealing with spoken language-based human-to-human communication. Once such a possibility is granted, the globalization of markets immediately requires that the related theme of multilinguality be also addressed. Clearly, if negotiation were important for e-commerce and e-service, the involved partners would appreciate, and actually require, that it be carried on in their own mother tongue. This cannot be achieved by having an e-commerce/service provider making available a set of human operators covering the relevant languages. That would not only be too expensive; it would also prove unfeasible in such domain as help-desks for trouble shooting and reparation, where the relevant operator should couple technical with language competence. On the other hand, the choice of using a given language as an 'interlingua', no matter how widely spread such as English, would irremediably exclude a large number of potential customers, especially in a non technical domain, as, e.g.. tourism.
Working on a class of applications in the sector of tourism, follows general consideration about the tendencies and needs of the international markets. Recent studies show that tourism is one of the economical sectors with more growth chances, and one which is going to be most affected by the availability of tele-

communication supports such as the web, video-conference, etc. (cite IFIT). Also, tourism in presently involved in an important change consisting in a shift from the current broker-supported market (agencies, tour operators), to a situation in which the customer directly contacts and negotiates with a local representative of local service providers the so called destinations. Such a change is important, since it is motivated by, and at the same time induces a parallel shift from ways to organize the offer around a fixed number of solutions, towards scenarios in which the keyword is flexibility. On the provider side, this requires the capability of providing personalized solutions meeting the demands and needs of customers (families, young children, old people, etc.), allowing for a better planning and allocation of resources and tourism presence. The customer, in turn, will have the possibility of explaining his/her needs and expectations, and a true guidance towards finding optimal solutions. The destination, on the other hand, needs that the system be capable of assisting him/her in managing data-bases of pictures and images describing hotels availability, sporting resorts, cultural events, etc., as well as textual information of various kinds. All this will enable the destination to react to the customer choice by proposing optimal solutions.

Working on a class of applications in the sector of video help-desk answers to pressing needs of companies which sells their products in different countries. This requires that local retailers or technicians communicate with the motherhouse to solve installation problems, fix bugs, adapt systems to users needs, etc. Some companies already have their own call centers addressing these needs, usually telephone- or web-based. However they are limited in coverage by language difference, and constrained in scope by the impossibility of widely resorting to visually presented information. In our scenario the customer might, for instance, require assistance to overcome problems in the installation of a given product and interact with the service provider in his/her own language. Visual manuals, pictures, and animations can be used in the course of the interaction, possibly allowing the two parties to comment upon them, try different solutions and providing for better chances to converge towards the solution of the problems in shorter time. Moreover, facilities can be envisaged to form databases of frequently asked questions (FAQ), which the server can access to further speed up the process.

Acknowledgements

The author thanks the team working on STST at ITC-irst for the research and technological contribution to the project: Mauro Cettolo, Anna Corazza, Fabio Pianesi, Emanuele Pianta and Lucia Tovena.

References

[1] Angelini B., Cettolo M., Corazza A., Falavigna D., and Lazzari G. (1997). "Multilingual Person to Person Communication at IRST". In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Munich. Germany.

[2] Breiman L., Friedman J., Olshen R., and Stone C. (1984). "Classification and Regression Trees". Wadsworth Inc.

[3] Brugnara F., Cettolo M. (1995). "Improvements in Tree-based Language Model Representation". In Proceedings of the 4th European Conference on Speech Communication and Technology. Madrid. Spain.

[4] Cancedda N., Kamstrup G., Pianta E., and Pietrosanti E. (1997). "Sax: Generating hypertext from sadt models". In Proceedings of the Third Workshop on Applications of Natural Language to Information Systems, Vancouver. Canada.

[5] Cettolo M., Corazza A., and De Mori R. (1998). "Language Portability of a Speech Understanding System". Computer Speech and Language. 12:1-21.

[6] De Mori R. and Kuhn R. (1995). "The Application of Semantic Classification Trees to Natural Language Understanding". IEEE Transactions on Pattern Analysis and Machine Intelligence. 17(5):449-460.

[7] Federico M., Cettolo M., Brugnara F., and Antoniol G. (1995). "Language Modelling for Efficient Beam-Search". Computer Speech and Language, 9(4):353-379.

[8] Cettolo M., Falavigna D. (1998). "Automatic Detection of Semantic Boundaries based on Acoustic and Lexical Knowledge". In Proceedings of the International Conference on Spoken Language Processing, Sidney. Australia.

[9] Moore R.C. (1994). "Integration of Speech with Natural Language Understanding". In D. B. Roe and J. G. Wilpon, editors, Voice Communication between Human and Machines, pages 254-271. National Academy Press. Washington D.C., USA.

[10] De Mori R., editor (1998). "Spoken Dialogues with Computers". Academic Press. San Diego, CA.

[11] Pianesi F. and Tovena L.M. (1998). "Using the interchange format for encoding spoken dialogues". In Proceedings of AMTA SIG-IL Second Workshop on Interlinguas and Interlingual Approaches.

[12] Pianta E. and Tovena L.M.(1998). "Generating with flexible templates from C-STAR Interchange Format". Technical Report 9808-04, Istituto per la Ricerca Scientifica e Tecnologica. ITC-irst, 1998.

[13] Rabiner L.R. (1990). "A tutorial on hidden Markov models and selected applications in speech recognition". In A. Waibel and K.F. Lee, editors, Readings in Speech Recognition, pages 267-296. Morgan Kaufmann Publishers. San Mateo, CA, 1990.

[14] Corazza A.(1999). "An Inter-Domain Portable Approach to Interchange Format Construction", accepted for publication in Proceedings of the European

Conference on Speech Communication and Technology. Budapest. 1999.
[15] Cettolo M., Corazza A., Lazzari G., Pianesi F., Pianta E. and Tovena L.M.(1999). "A Speech-to-Speech Translation based Interface for Tourism". In Proceedings of the ENTER'99 Conference. Innsbruck. Austria.
[16] Lazzari G., editor (1998). "Speaker Language Identification and Speech Translation - Chapter 7". In Hovy E., Ide N., Frederking R., Mariani J., Zampolli A., editors. (1998). Multilingual Information Management: Current Levels and Future Abilities. A study commissioned by the US National Science Foundation. http://www.cs.cmu.edu/~ref/mlim/.