Speech Translation for French in the NESPOLE! European Project

L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud
S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz

CLIPS-IMAG Laboratory, Joseph Fourier University
B.P. 53, 38041 Grenoble cedex 9, France
Laurent.Besacier@imag.fr

Abstract

This paper presents CLIPS laboratory activities in the context of the NESPOLE! European project, exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors. The scientific and technological research issues particularly addressed in order to improve current experimental speech-to-speech translation systems, are: robustness, scalability, and cross-domain portability. The general architecture of the whole speech to speech translation demonstrator is first presented and the Interchange Format (IF) strategy for translation adopted in the project is quickly described. The French database recorded during the project and the French Human Language Technology (HLT) modules (recognition, synthesis and translation) are then fully detailed. First results obtained and future perspectives of the project are also discussed in this article.

1. Introduction

The NESPOLE![1] Project is a common EU NSF funded project exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors [1]. The spoken languages involved in this project are Italian, German, English and French. Partners of the project are *ITC/IRST* from Trento (Italy), *ISL Labs*. from *UKA* (Karlsruhe, Germany) and *CMU* (Pittsburgh, USA), *Aethra* (an italian company specialised in videoconferencing software), *APT* : a tourism agency in the Trentino area (Italy) and finally CLIPS laboratory (Grenoble, France).

The scenario for the first showcase of NESPOLE! involves an Italian speaking agent, located in a tourism agency in Italy and a client located anywhere (English, German or French speaking) using a simple terminal (PC, sound and video cards, H323 videoconferencing software like NetMeeting™). This choice is related to present available technology, in the near future the third generation cellular can be also used as terminal.

The client wants to organise a trip in the Trentino area, and refers to APT (the tourism agency) web pages in order to get information. If the client wants to know more about a particular topic or prefers to have a more direct contact, a speech to speech translation service allows him to interact in his own language with an APT Italian agent. A videoconferencing session can then be opened between client and agent and the dialog starts between them. The scenario starts from the assumption that the tourist has already visited the APT site *www.trentino.to*. However, as usually happens and as a brief analysis done on e-mail received by APT related to general information request confirms, the tourist has not searched the web site nor read the pages in detail and wants to ask some details about a specific subject.

In this project, the scientific and technological research issues intended to be addressed in order to improve current experimental speech-to-speech translation (STST) systems, are: robustness, scalability, cross-domain portability and multimodal interaction with multimedia content.

This paper first describes the general architecture of the whole speech to speech translation demonstrator. The French database recorded during the project and the French Human Language Technology (HLT) modules (recognition, synthesis and translation) are then detailed in this article.

2. Architecture of the system

2.1. IF (Interchange Format) strategy for translation

An Interchange Format (IF) based approach was adopted in the project. IF 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.

The IF [2] relies on dialogue acts, concepts, and arguments. Dialogue acts describe speaker's intention, goal, and need**.** Concepts define the focus of the dialogue act. Several concepts may appear in one IF. Arguments are values of discourse variables.

The global architecture for speech translation using the IF approach is thus the one described in *Figure 1*:

[1] see http://nespole.itc.it/

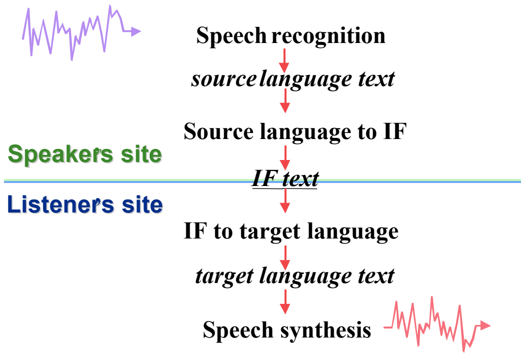

Figure 1: Overall components interaction

2.2. Overall NESPOLE architecture

This section presents an overview of the hardware and software infrastructure for Nespole! (see *figure 2*). The design is based on the experience carried on in the C-STAR II[2] Consortium and taking into account the present technology (H323) and, with minor changes, future developments (UMTS for example). The overall architecture has been designed taking into account the geographical location of the four Language Specific HLT Servers, and assuming complete structural symmetry of Agents and Clients.

Moreover, monitoring tasks have been distributed among four distinct hosts. Agent and Client are H.323 terminals namely, they are able to make H.323 calls, sending audio in G.711 (or G72X) format, video in H.261 format, and data in T.120 format. The Global Nespole! Server is drawn as a unique entity, but the various Language Specific HLT Servers can be arbitrarily distributed.

The Mediator Software Module is in charge of interfacing the language specific HLT servers and the clients. All the communications among Agent, Client, Mediator and HLT Server are via sockets. This is particularly appropriate for communications between Mediator and HLT Server, since transmission via sockets is independent from the actual operating systems running at the endpoints. Communications between Agent and Mediator, and between Client and Mediator are ruled by the H.323 standard.

Assuming that the Client speaks a language X and wants to communicate with an Agent speaking language Y, we denote by MY and SY the Mediator and HLT Server for language Y. Analogous abbreviations hold for MX, SX, etc.

Client makes a H.323 call to MY. The call mode allows the Mediator to identify the source language X. The Mediator, which acts as a bridge between Client and Agent, makes a H.323 call to the Agent for language Y. Sockets for audio, video, and data packets are opened between Client and MY, and between MY and Agent, respectively. The mediator is also connected to the HLT servers of the agent's and client's language and send / receive adequate data to / from the HLT modules.

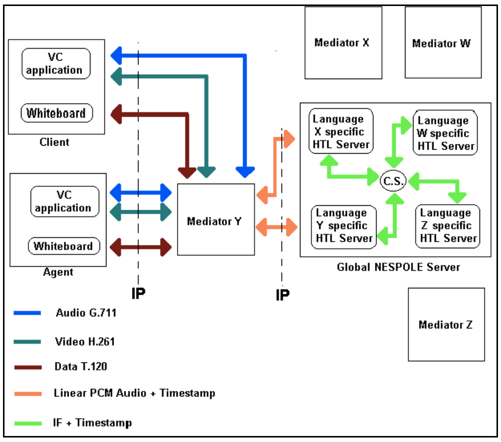

Figure 2 : Architecture in Nespole

3. French data collection

For IF specification and vocabulary definition, a large data collection was performed in summer 2000 in each partners language [3].

191 dialogs in 4 languages were recorded, involving a tourist client that asks about information to an agent located in APT (tourism agency) in Italy. 5 different scenarios were developed (winter accommodation, all-included tourist package, summer vacation in a park, castle and lake tours, and looking for folklore and brochures).

31 French dialogs were recorded using the videoconferencing software NetMeeting. Speech was recorded on both sites (agent and client site). Thus, on the agent site, local speech from the agent and VoIP (Voice over IP) speech from the client were collected, whereas on the client site, local speech from the client and VoIP speech from the agent were collected. The G723.1 speech coder was selected during this data collection

Manual transcriptions of the 31 recorded French dialogs were used to determine a task vocabulary (related to the scenarios defined) of 2056 words.

4. Speech to speech translation components

GEOD[3] team of CLIPS Laboratory developed the French speech recogniser and integrated a French speech synthesiser in the demonstrator, whereas GETA[4] team of CLIPS developed the French to IF analyser and the IF to French generator.

4.1. Speech Recognition

Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [4] from CMU. The context dependent acoustic model was learned on a corpus which contains 12 hours of continuous speech of 72 speakers extracted from Bref80 [5] corpus. The French speech recognition module, which is an essential part of the complete speech to speech translation chain, was adapted in two ways. First, the language model had to deal with task specific vocabulary and secondly, the acoustic model had to deal with VoIP speech.

[2] http://www.c-star.org/

[3] http://www-clips.imag.fr/geod/

[4] http://www-clips.imag.fr/geta/

4.1.1. Language model adaptation

In a former paper [6] we have shown that Internet documents can increase word accuracy by giving very large amount of training data for spoken language modelling. The base of our method, called “minimal blocks”, is the vocabulary. Therefore, the first step for the language model adaptation phase is the definition of the task vocabulary which was done during data collection phase (see *section 3*). In parallel, we also computed a word frequency count on *WebFr4* (our last collection of French Web pages, a few less than 6 millions documents, more than 44 Go of data). This list contains about 200,000 different French inflected forms. The most frequent forms were used to add general words to our domain dependent vocabulary. Final dictionary is 20,000 lexical forms long.

The second step is language model training. To do that, we used our minimal block method with the vocabulary obtained in the first step. From *WebFr4* a training set of 1,528,347,218 words was extracted. We computed (using our adapted tools) a trigram language model on these data. Due to the amount of data used for training phase, cut-offs of language model events, were set to higher values than usually. In our experiments, we set them to 10 and 100 for bigrams and trigrams respectively.

4.1.2. Acoustic model adaptation

In NESPOLE! scenario, a client can use the speech translation service from anywhere. Thus, speech needs to be achieved from the client terminal to the speech to speech translation modules, the first module being speech recognition. When a videoconferencing software is used, the speech signal is coded to reduce transmission delays (G711, G723.1, G728 codecs…). As transcoding (the process of coding and decoding) modifies the speech signal, it is likely to have an influence on speech recognition performance if acoustic models are not adapted. The solution we propose is to train the acoustic models with transcoded speech. For this, the BREF80 database (that we use for training our acoustic models) was transcoded via different coders in order to train one acoustic model for each speech coder used in our application. To know which speech coder is used on the user terminal (and thus to know which acoustic model must be used) the information concerning the coder used by the client is encoded in the data stream transmitted from the client to the speech recognition server.

4.1.3. Performance

To evaluate the speech recognition performance we tried to recognise 77 sentences transmitted on IP network from the client to the French HLT server (sentences taken from the transcription of French dialogs); the word error rate obtained was 18%.

4.2. French-to-IF analyser

4.2.1. Context

The ARIANE-G5 environment [7], in which each of the components is a rewriting rules system, imposes *de facto* a linguistic-based MT approach [8]. Our partners have preferred a stochastic-based approach they find more suitable to handle syntax-relaxed input.

The transcribed oral inputs to the ARIANE process are thus syntax-relaxed (the grammar of spoken utterances is not the grammar of written ones). Moreover, speech recognition errors combined with the intrinsic ambiguity of any natural language (homophones) imposes us to deal with ill-formed input as compared with the written and even the clean oral syntax.

Therefore, we gave up with the idea of a complete syntactical analysis of the input and chose to recognize structured fragments relevant for the IF (phrases describing dates, prices, localization, quantities, …). The other fragments, incomplete and isolated, according to their content, are word spotted or simply ignored and removed.

4.2.2. Implementation

ARIANE receives a textual transcription of the spoken utterance in the source language and outputs a text in the target language (IF). From the source language to the target language we use a succession of module pairs consisting in grammars and dictionaries. They make the analysis of the source language (at different levels) and the translation of the lexical items towards the target language and finally, the generation of the output which take into account the distributional properties of the target vocabulary. The production steps of the IF are the following ones:

- Morphological analysis and lemmatization of the words of the input;
- First dictionary look-up to prepare the analysis and the transfer;
- Syntactical analysis of the semantically relevant fragments : dates, quantities, numbers, prices , etc.;
- Transfer dictionaries look-up towards the IF;
- Syntactical IF generation (planning, ordering of the constituents within the IF syntax) and finally morphological generation of the output.

4.2.3. Evaluation

Starting from C-STAR II, the IF definition went through a huge amount of changes for Nespole!. The IF is now meant to describe far more detailed information. The rough techniques used for C-STAR II are being refined to fit the new needs of Nespole project.

4.3. IF-to-French generator

4.3.1. Context

Within C-STAR II, the CLIPS laboratory was not in charge of the generation from IF [9]. This is a new task we are dealing with.

We have decided to use the specification of the IF as a start-up point to feed the IF-to-French module. All the IF specification files were automatically parsed to produce as much as possible data automatically, that is the non-terminals and the terminals of the grammars, the continuations (see infra). Problems encountered were due to the fact that the current IF specification is not "clean" enough for all this process to be done fully automatically.

The advantage of our methodology is to reach a full coverage of all the possible realizations of an IF. The drawback is that the development process is slowed down at the beginning due to the necessity of describing "linguistically" every terminal item of the IF.

4.3.2. Principle

The specification of the IF gives for each item (speech act, concept, argument or value) its possible right hand-side continuation. We have semantically labeled each link between a terminal symbol of the IF with all its possible continuations. The top-down recognition of an IF input consists in the following steps:

- associate to each terminal of the IF input the list of its possible labelled continuations;
- built a semantic tree of the IF according to the continuations;
- translate all the terminals and structures of the IF into French words and structures;
- generate a syntactically well-formed sentence in French.

4.3.3. Implementation

ARIANE receives an IF and outputs a French language text. An IF tree is then transformed into a French linguistic tree that will be used as the input of a general purpose French generation module towards written output. It is planned to modify this generator to produce speech-oriented outputs. The production steps of the French output are thus the following ones:

- splitting of the IF text into the IF terminals and construction of a flat tree;
- dictionary look-up for the continuations;
- construction of a labelled tree with the effective continuations;
- lexical transfer from IF to French;
- structural transfer from an IF semantic tree to a French logic and semantic tree;
- syntactic and morphological generation of a French text to be synthesised.

4.4. French speech synthesiser

The last part of automatic translation process is the synthesis. Text-to-speech (TTS) synthesis involves the computation of a speech signal from input text.

We chose to implement the synthesis as an internet server that will receive text from the user and will respond by sending back a sound file. In order to minimise the time needed for the entire translation process the synthesis must be done as fast as possible.

The Euler TTS[5] [10] system of the Polytechnical Faculty of Mons was considered fast enough for our project. It needs about 30% of the sentence time length to produce the corresponding sound. The quality of the synthetic speech is acceptable for our application.

5. Conclusions and Perspectives

We have presented the context of the NESPOLE! Project, and the speech to speech translation modules for French developed at CLIPS laboratory for this project.

The choice of the Interchange Format (IF) strategy for translation allows us to envisage new perspectives like, for instance, the encoding in the IF of prosodic marks extracted by the speech recogniser. Then, these tags could be used to generate best prosody and try to reproduce original speaker intonation with the speech synthesiser. IF could also convey multimodal information like gestures (pointing, selection or free hand strokes on a whiteboard). This multimodal content is planned to be addressed in the next developments of NESPOLE project.

Finally, to decisively improve the quality and usability of speech to speech translation systems, considering the dialogue context [11], i.e. having a representation of the past dialog, is an interesting issue.

6. References

[1] Lazzari G., Spoken translation : challenges and opportunities, *ICSLP'2000*, Beijing, China. Oct. 16-20, 2000, vol 4/4 : pp. 430-435

[2] Levin L. & al. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. *Proc. ICSLP'98*, 30th November - 4th December 1998, Sydney, Australia, vol.4/7, pp.1155- 1158.

[3] Burger, S., Besacier, L. Metze, F., Morel, C., Coletti, P., The NESPOLE! VoIP dialog database, submitted to *Eurospeech 2001.*

[4] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS : A Speech Translation System". *Eurospeech*, 1993, volume 2, pages 1295-1298.

[5] Lamel, L.F., Gauvain, J.L., Eskénazi, M. "BREF, a Large Vocabulary Spoken Corpus for French", *Eurospeech*, Gênes, Italy, Vol 2, pp. 505-508, 24-26 September 1991.

[6] Vaufreydaz, D., Akbar, M., Rouillard, J., Internet documents : a rich source for spoken language modeling, *ASRU'99 Workshop*, Keystone Colorado (USA), pp. 277-280.

[7] Boitet C. 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 September 1997, Meisei University, vol. 1/1 : pp. 23-57.

[8] Boitet C. & Guilbaud J.-P. Analysis into a formal task-oriented pivot without clear abstract semantic is best handled as usual translation. *Proc. ICSLP-2000*, Oct. 16-20, 2000, vol 4/4 : pp. 436-439.

[9] Blanchon, H. & Boitet, C. (2000). Speech Translation for French within the C-STAR II Consortium and Future Perspectives. *Proc. ICSLP 2000*. Beijing, China. 16-20 October, 2000. vol. 4/4 : pp. 412-417.

[10] Bagein, M., Dutoit, T., Malfrere, F., Pagel, V., Ruelle, A., Tounsi, N., Wynsberghe, D. "An Open, Generic, Multi-lingual and Multi-platform Text-to-Speech System".

[11] Boitet, C., Blanchon, H, Guilbaud J.-P. (2000). A way to integrate context processing in the MT component of spoken, task-oriented translation systems. Proc. MSC2000. Kyoto, Japon. October 11-13, 2000. vol. 1/1 : pp. 83-87.

[5] http://tcts.fpms.ac.be/synthesis/euler