Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Applications

Alon Lavie,
Chad Langley,
Alex Waibel
Carnegie Mellon University
Pittsburgh, PA, USA
alavie@cs.cmu.edu

Fabio Pianesi,
Gianni Lazzari,
Paolo Coletti
ITC-irst
Trento, Italy

Loredana Taddei,
Franco Balducci
AETHRA
Ancona, Italy

1. INTRODUCTION

NESPOLE! [1] is a speech-to-speech machine translation research project funded jointly by the European Commission and the US NSF. The main goal of the NESPOLE! project is to advance the state-of-the-art of speech-to-speech translation in a real-world setting of common users involved in e-commerce applications. The project is a collaboration between three European research labs (IRST in Trento Italy, ISL at University of Karlsruhe in Germany, CLIPS at UJF in Grenoble France), a US research group (ISL at Carnegie Mellon in Pittsburgh) and two industrial partners (APT - the Trentino provincial tourism bureau, and Aethra - an Italian tele-communications commercial company). The speech-to-speech translation approach taken by the project builds upon previous work that the research partners conducted within the context of the C-STAR consortium (see `http://www.c-star.org`). The prototype system developed in NESPOLE! is intended to provide effective multi-lingual speech-to-speech communication between all pairs of four languages (Italian, German, French and English) within broad, but yet restricted domains. The first showcase currently under development is in the domain of tourism and travel information.

The NESPOLE! speech translation system is designed to be an integral part of advanced e-commerce technology of the next generation. We envision a technological scenario in which multi-modal (speech, video and gesture) interaction plays a significant role, in addition to the passive browsing of pre-designed web pages as is common in e-commerce today. The interaction between client and provider will need to support online communication with agents (both real and artificial) on the provider side. The language barrier then becomes a significant obstacle for such online communication between the two parties, when they do not speak a common language. Within the tourism and travel domain, one can imagine a scenario in which users (the clients) are planning a recreational trip and are searching for specific detailed information about the regions they wish to visit. Initial general information is obtained from a web site of a tourism information provider. When more detailed or special information is required, the customer has the option of opening an online video-conferencing connection with a human agent of the tourism information provider. Speech translation is integrated within the video-conference connection; the two parties each speak in their native language and hear the synthesized translation of the speech of the other participant. Text translation (in the form of subtitles) can also be provided. Some multi-modal communication between the parties is also available. The provider agent can send web pages to the display of the customer, and both sides can annotate and refer to pictures and diagrams presented on a shared whiteboard application.

In this paper we describe the design considerations behind the architecture that we have developed for the NESPOLE! speech translation system in the scenario described above. In order to make the developed prototype as realistic as possible for use by a common user, we assume only minimal hardware and software is available on the customer side. This does include a PC-type video camera, commercially available internet video-conferencing software (such as Microsoft Netmeeting), standard audio and video hardware and a standard web browser. However, no speech recognition and/or translation software is assumed to reside locally on the PC of the customer. This implies a server-type architecture in which speech recognition and translation are accomplished via interaction with a dedicated server. The extent to which this server is centralized or distributed is one of the major design considerations taken into account in our system.

2. NESPOLE! INTERLINGUA-BASED TRANSLATION APPROACH

Our translation approach builds upon previous work that we have conducted within the context of the C-STAR consortium. We use an interlingua-based approach with a relatively shallow task-oriented interlingua representation [2] [1], that was initially designed for the C-STAR consortium and has been significantly extended for the NESPOLE! project. Interlingual machine translation is convenient when more than two languages are involved because it does not require each language to be connected by a set of transfer rules to each other language in each direction [3]. Adding a new language that has all-ways translation with existing languages requires only writing one analyzer that maps utterances into the interlingua and one generator that maps interlingua representations into sentences. The interlingua approach also allows each partner group to implement an analyzer and generator for its home language only. A fur-

[1] NESPOLE! - NEgotiating through SPOken Language in E-commerce. See the project website at `http://nespole.itc.it/`

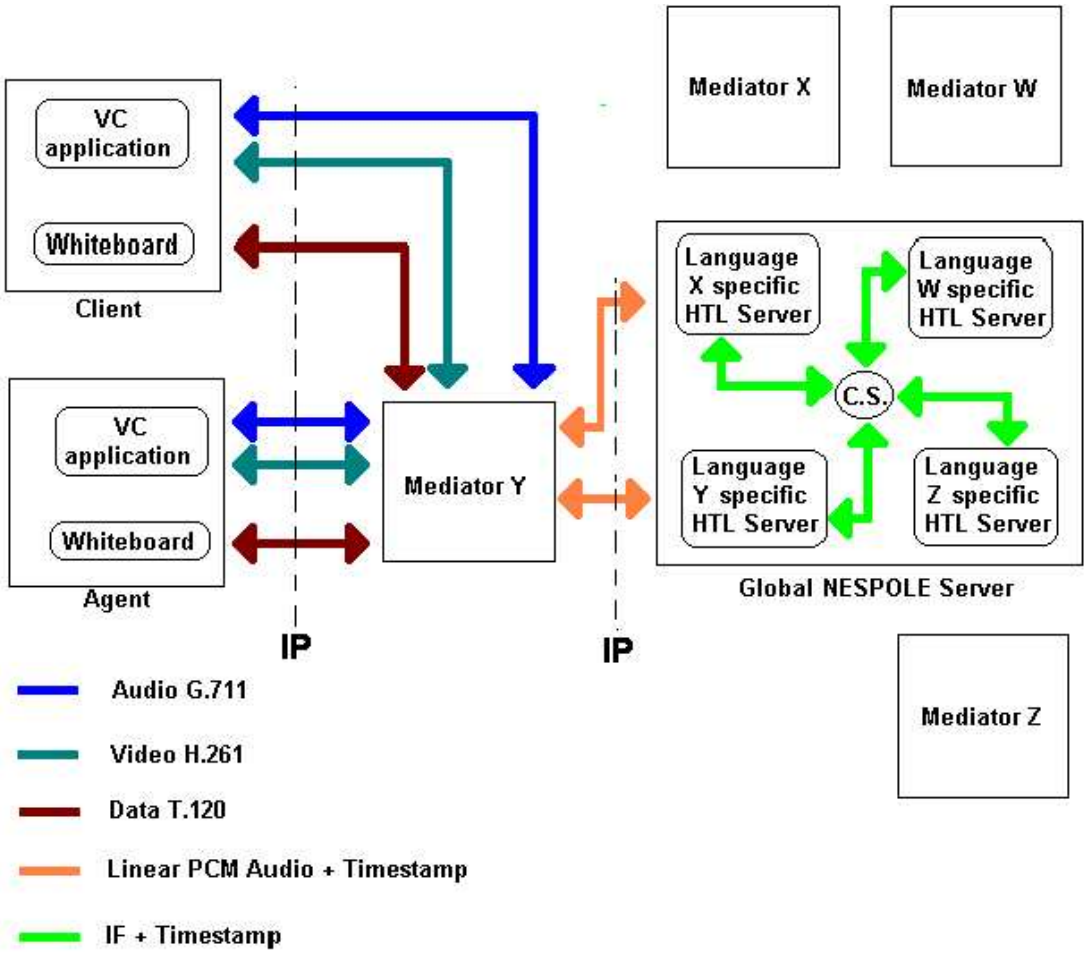

Figure 1: General Architecture of NESPOLE! System

ther advantage is that it supports a paraphrase generation back into the language of the speaker. This provides the user with some control in case the analysis of an utterance failed to produce a correct interlingua. The following are three examples of utterances tagged with their corresponding interlingua representation:

```
Thank you very much
c:thank

And we'll see you on February twelfth.
a:closing (time=(february, md12))

On the twelfth we have a single and a double
available.
a:give-information+availability+room
   (room-type=(single & double),time=(md12))
```

3. NESPOLE! SYSTEM ARCHITECTURE DESIGN

Several main considerations were taken into account in the design of the NESPOLE! Human Language Technology (HLT) server architecture: (1) The desire to cleanly separate the actual HLT system from the communication channel between the two parties, which makes use of the speech translation capabilities provided by the HLT system; (2) The desire to allow each research site to independently develop its language specific analysis and generation modules, and to allow each site to easily integrate new and improved components into the global NESPOLE! HLT system; and (3) The desire of the research partners to build to whatever extent possible upon software components previously developed in the context of the C-STAR consortium. We will discuss the extent to which the designed architecture achieves these goals after presenting an overview of the architecture itself.

Figure 1 shows the general architecture of the current NESPOLE! system. Communication between the client and agent is facilitated by a dedicated module - the *Mediator*. This module is designed to control the video-conferencing connection between the client and the agent, and to integrate the speech translation services into the communication. The mediator handles audio and video data associated with the video-conferencing application and binary data associated with a shared whiteboard application. Standard H.323 data formats are used for these three types of data transfer. Speech-to-speech translation of the utterances captured by the mediator is accomplished through communication with the NESPOLE! global HLT server. This is accomplished via socket connections with language-specific HLT servers. The communication between the mediator and each HLT server consists mainly of linear PCM audio packets (some text and control messages are also supported and are described later in this section).

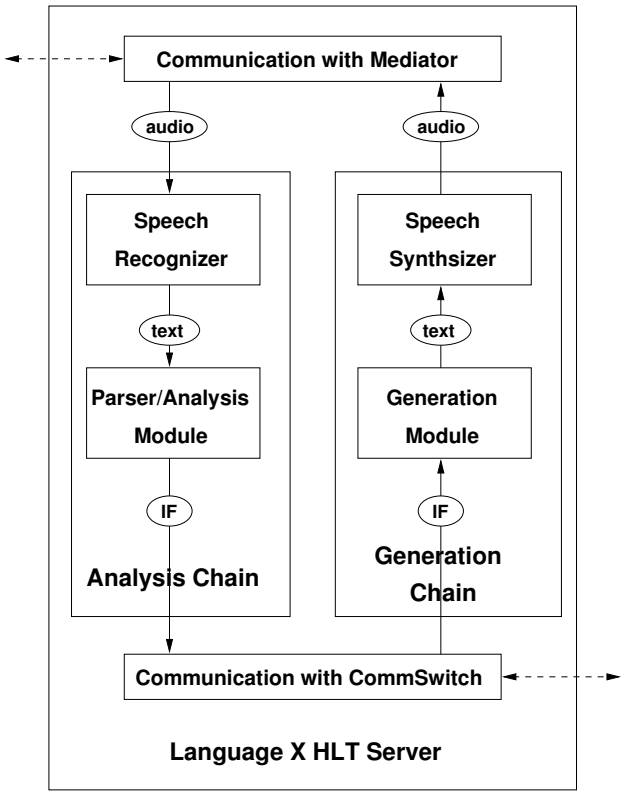

Figure 2: Architecture of NESPOLE! Language-specific HLT Servers

The global NESPOLE! HLT server comprises four separate language-specific servers. Additional language-specific HLT servers can easily be integrated in the future. The internal architecture of each language-specific HLT server is shown in figure 2. Each language-specific HLT server consists of an *analysis chain* and a *generation chain*. The analysis chain receives an audio stream corresponding to a single utterance and performs speech recognition followed by parsing and analysis of the input utterance into the interlingua representation (IF). The interlingua is then transmitted to a central HLT communication switch (the CS), that forwards it to the HLT servers for the other languages as appropriate. IF messages received from the central communication switch are processed by the generation chain. A generation module first generates text in the target language from the IF. The text utterance is then sent to a speech synthesis module that produces an audio stream for the utterance. The audio is then communicated externally to the mediator, in order to be integrated back into the video-conferencing stream between the two parties.

The mediator can, in principle, support multiple one-to-one communication sessions between client and agent. However, the design supports multiple mediators, which, for example, could each be dedicated to a different provider application. Communication with the mediator is initiated by the client by an explicit action via the web browser. This opens a communication channel to the mediator, which contacts the agent station, establishes the video-conferencing connection between client and agent, and starts the whiteboard application. The specific pair of languages for a dialogue is determined in advance from the web page from which the client initiates the communication. The mediator then establishes a socket communication channel with the two appropriate language specific HLT servers. Communication between the two language specific HLT servers, in the form of IF messages, is facilitated by the NESPOLE! global communication switch (the CS). The language specific HLT servers may in fact be physically distributed over the internet. Each language specific HLT server is set to service analysis requests coming from the mediator side, and generation requests arriving from the CS.

Some further functionality beyond that described above is also supported. As described earlier, the ability to produce a textual paraphrase of an input utterance and to display it back to the original speaker provides useful user control in the case of translation failures. This is supported in our system in the following way. In addition to the translated audio, each HLT server also forwards the generated text in the output language to the mediator, which then displays the text on a dedicated application window on the PC of the target user. Additionally, at the end of the processing of an input utterance by the analysis chain of an HLT server, the resulting IF is passed internally to the generation chain, which produces a text generation from the IF. The result is a textual paraphrase of the input utterance in the source language. This text is then sent back to the mediator, which forwards it to the party from which the utterance originated. The paraphrase is then displayed to the original speaker in the dedicated application window. If the paraphrase is wrong, it is likely that the produced IF was incorrect, and thus the translation would also be wrong. The user may then use a button on the application interface to signal that the last displayed paraphrase was wrong. This action triggers a message that is forwarded by the mediator to the other party, indicating that the last displayed translation should be ignored. Further functionality is planned to support synchronization between multi-modal events on the whiteboard and their corresponding speech actions. As these are in very preliminary stages of planning we do not describe them here.

4. DISCUSSION AND CONCLUSIONS

We believe that the architectural design described above has several strengths and advantages. The clean separation of the HLT server dedicated to the speech translation services from the external communication modules between the two parties allows the research partners to develop the HLT modules with a large degree of independence. Furthermore, this separation will allow us in the future to explore other types of mediators for different types of applications. One such application being proposed for development within the C-STAR consortium is a speech-to-speech translation service over mobile phones. The HLT server architecture described here would be able to generally support such alternative external communication modalities as well.

The physical distribution of the individual language specific HLT servers allows each site to independently develop, integrate and test its own analysis and generation modules. The organization of each language specific HLT server as an independent module allows each of the research sites to develop its unique approaches to analysis and generation, while adhering to a simple communication protocol between the HLT servers and externally with the mediator. This allowed the research partners to "jump-start" the project with analysis and generation modules previously developed for the C-STAR consortium, and incrementally develop these modules over time. Furthermore, the global NESPOLE! communication switch (the CS) supports testing of analysis and generation among the four languages in isolation from the external parts of the system. Currently, requests for analysis of a textual utterance can be transmitted to the HLT servers via the CS, with the resulting IF sent (via the CS) to all HLT servers for generation. This gives us great flexibility in developing and testing our translation system. The functionality of the CS was originally developed for our previous C-STAR project, and was reused with little modification.

Support for additional languages is also very easy to incorporate into the system by adding new language-specific HLT servers. Any new language specific HLT server needs only to adhere to the communication protocols with both the global NESPOLE! communication switch (the CS) and the external mediator. The C-STAR consortium plans to use the general architecture described here for its next phase of collaboration, with support for at least three asian languages (Japanese, Korean and Chinese) in addition to the languages currently covered by the NESPOLE! project.

The first prototype of the NESPOLE! speech translation system is currently in advanced stages of full integration. A showcase demonstration of the prototype system to the European Commission is currently scheduled for late April 2001.

5. ACKNOWLEDGMENTS

The research work reported here was supported in part by the National Science Foundation under Grant number 9982227. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

6. REFERENCES

[1] L. Levin, D. Gates, A. Lavie, F. Pianesi, D. Wallace, T. Watanabe, and M. Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogue. In *Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, Seattle, 2000.

[2] L. Levin, D. Gates, A. Lavie, and A. Waibel. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, pages Vol. 4, 1155–1158, Sydney, Australia, 1998.

[3] S. Nirenburg, J. Carbonell, M. Tomita, and K. Goodman. *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann, San Mateo, California, 1992.