The NESPOLE! Multimodal Interface for Cross-lingual Communication - Experience and Lessons Learned

Loredana Taddei
Aethra Telecomunicazioni
Ancona, Italy
l.taddei@aethra.it

Erica Costantini
Department of Psychology
University of Trieste, Italy
costanti@psico.units.it

Alon Lavie
Carnegie Mellon University
Pittsburgh, PA, USA
alavie@cs.cmu.edu

Abstract

In this paper we describe the design, evolution, and development of the user interface components of the NESPOLE! speech-to-speech translation system. The NESPOLE! system was designed for users with medium-to-low levels of computer literacy and web expertise. The user interface was designed to effectively combine web browsing, real-time sharing of graphical information and multi-modal annotations using a shared whiteboard, and real-time multilingual speech communication, all within an e-commerce scenario. Data collected in sessions with naïve users in several stages in the process of system development formed the basis for improving the effectiveness and usability of the system. We describe this development process, the resulting interface components and the lessons learned.

1. Introduction

The demands for user-friendly interfaces to interactive computer applications have significantly grown in recent years. In the development of software applications, a large amount of time is dedicated to the study and realization of the program interface, which had often been neglected in the past. The improvement of graphic tools available to developers and the increase in computation power of computers have contributed to the growth of interest in user interfaces. Moreover, it has become common practice to carry out psychological studies for investigating the user's cognitive skills and their preferences and expectations concerning interfaces.

NESPOLE! (Negotiation through SPOken Language in E-commerce) is a speech-to-speech machine translation project designed to provide fully functional speech-to-speech capabilities within real-world settings of common users involved in e-commerce applications. The project is a collaboration between three European research groups (IRST in Trento, Italy; ISL at Universität Karlsruhe (TH); and CLIPS at Université Joseph Fourier in Grenoble, France), one US research group (ISL at Carnegie Mellon University in Pittsburgh, PA) and two industrial partners (APT; Trento, Italy – the Trentino provincial tourism board, and AETHRA Telecomunicazioni; Ancona, Italy – a telecommunication company). The project is funded jointly by the European Commission and the US NSF.

The first NESPOLE! showcase was developed for the following scenario: while navigating on the Internet page of the Tourism Board of Trentino (an Italian region) a French, English or German user is interested in additional information that is not contained on the web pages. Using a dedicated link from the provider web page, the user can establish a live real-time interactive connection with an Italian human operator of the Tourism Board. The system's interactive multimodal applications are then automatically activated and support the communication between the two parties.

2. The NESPOLE! System

The design principles of the NESPOLE! system are described in [1]. The NESPOLE! system uses a client-server architecture to allow a common user, who is initially browsing through the web pages of a service provider on the Internet, to connect seamlessly to a human agent of the service provider who speaks another language, and provides speech-to-speech translation service between the two parties. Standard commercially available PC video-conferencing technology such as Microsoft® NetMeeting® is used to connect between the two parties in real-time.

A key component in the NESPOLE! system is the "Mediator" module, which is responsible for mediating

the communication channel between the two parties as well as interfacing with the appropriate Human Language Technology (HLT) speech-translation servers (Figure 1). The HLT servers provide the actual speech recognition and translation capabilities.

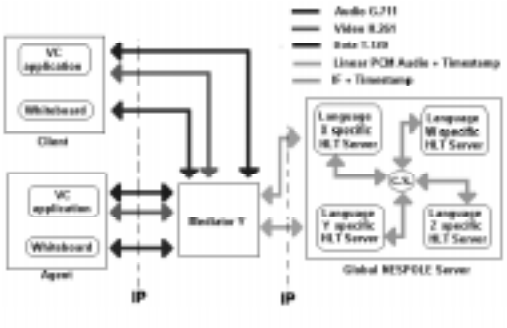

Figure 1. NESPOLE! System Architecture

This system design allows for a very flexible and distributed architecture: Mediators and HLT-servers can be run in various physical locations, so that the optimal configuration (including fall-back strategies) given the locations of the client and the agent and anticipated network traffic can be taken into account at any time. A well-defined API allows the HLT servers to communicate with each other and with the Mediator, although the HLT modules within the servers for the different languages are implemented using very different software packages.

The modularity of the NESPOLE! system design has provided us with flexibility in the independent development and testing of various components and modules of the system. For example, early in the project, we were able to test the Global NESPOLE! Server without the H.323 connections, and consequently without the complications of the audio coding and decoding problems linked to these connections. This allowed us to isolate individual problems and to cope with one problem at a time. Furthermore, during pre-tests prior to the multimodality experiment (see § 5.1), we were able to run the system with translation disabled, allowing us to record monolingual dialogs, during which we could easily evaluate several system features (e.g. the functions of the AeWhiteboard and the pen-tablet device).

3. User interface

The NESPOLE! user interface at both Client and Agent sites is composed of 4 windows: a web browser, Microsoft® NetMeeting®, the AeWhiteboard and the NESPOLE! Monitor. Each window plays a different role in the interaction between client and agent and their activation is automatic: the user does not need to worry about opening or positioning them on the desktop screen because these operations are carried out by the system in order to optimize space and usability.

3.1 Activation of the System

The user can start a videoconference with the Italian operator of the Tourism Board by simply pressing a dedicated hyperlink on the web page. The following events then take place:

- activation of Microsoft® Netmeeting®,
- establishment of the audio-video-data call to the system of the Tourism Board operator in Trentino,
- transmission of the user's web page address to the agent; this allows the browser on the agent's PC to display the exact same page as the one seen by the user; moreover, if the web page is available in different languages, the agent PC will display the Italian page that corresponds to the French, English or German page of the user;
- activation of AeWhiteboard for graphic information exchange;
- activation of NESPOLE! Monitor to keep track of the translation process provided by the Global NESPOLE! translation server.

These functions have been implemented by using NetMeeting® UI ActiveX Control from Microsoft® and ApplLaunch.ocx developed by Aethra.

3.2 Microsoft® NetMeeting®

The establishment of the audio-video-data call with the Tourism Board operator in Trentino allows the user to view the video images of the agent "live" in one of Netmeeting® areas. The two users can then communicate, each speaking their own language thanks to the audio decoding, voice recognition, analysis, synthesis and new coding implemented by the Mediator and the Global NESPOLE! Server.

Each user is able to hear both the original audio from the remote user as well as the translation of this audio as provided by the system. The two audio streams are mixed and can overlap. This functionality, provided and managed by Mediator modules, simulates the "simultaneous" translation capabilities that would be provided by a human interpreter. In our case, where network traffic and translation processes introduce time delays, the ability to hear the original audio provides the users with appropriate feed-back on what is taking place on the other side (the partner is waiting or the partner is speaking). This feedback is very useful to avoid overlapping speech, especially in case that video transmission is not available (see § 5.2).

The interface controlling the Mediator supports the disabling of original audio transmission and controls

both original and translated audio volume independently. This feature is useful in case there is a need for more controlled situations. For example, during one of our experiments [3, 4] we disabled the original audio in order to ensure that verbal information was being communicated *only* via the translation (and not via the original language).

NetMeeting® delivers additional functions to make the exchange of information and communication easier: audio volume control on the user side and the possibility of muting the local audio. These functions are especially useful in the case of very noisy environments. Moreover, the data channel opened by NetMeeting is in compliance with the T.120 standard, which allows for file transfer and application sharing.

3.3 The AeWhiteboard

The AeWhiteboard (Figure 2) is based on Windows application standards and features menus, a tool bar and a status bar. It allows the user to view bitmaps, such as, in the current NESPOLE! showcase, town maps or maps of tourist areas in Trentino. The user can also draw gestures on the bitmap to show routes or highlight places, zoom in and out and scroll the bitmap.

Figure 2. The AeWhiteboard

The AeWhiteboard drawing functionalities include:

- free-hand strokes: the user can draw arrows, lines, circles and other free-hand stokes of choice;
- lines: the user can connect two points on the maps by a line;
- selection of areas on the map: this can be done by enclosing any area on the map within an elliptical/rectangular figure.

The drawings are performed by means of a tablet-pen device. Appropriate colors can be selected among the palette for all types of drawings, to distinguish among different gestures. Finally, the user can save a copy of the map with the drawings performed on it, in order to reload it whenever needed.

What makes this application extremely useful for communication is the fact that the user can share operations and drawings with the remote user. The results of an experiment [3, 4] and the behavior of the users who interacted through the NESPOLE! system in several occasions (see § 5.1) suggest that visual information, and in particular gestures and drawings, can dramatically improve dialogue effectiveness. The ability to complement speech with such non-verbal exchanges of information often helps users resolve misunderstandings and ambiguities that are due to recognition and translation errors.

Another important functionality supported by the AeWhiteboard is the ability to simultaneously display a web page on the browsers of both parties. If the same page is available in multiple languages, the system will display the web page in the appropriate language of each of the two users.

All of the above tools and modalities are available to both parties throughout the communication, interleaved with the ongoing multilingual verbal dialogue that is taking place. The goal is to allow the two users to act and feel as if they were sitting around a table exchanging brochures and illustrative material.

Data exchange between the two AeWhiteboard applications and Mediator takes place by the means of a proprietary protocol developed in the T.120 standard communication channel within the H.323 audio-video-data connection over IP networks.

The AeWhiteboard implements a gesture recognition mechanism based on the idea that strokes of the pen within time intervals of less than 1.5 seconds can be considered to be part of the same gesture. For example, the user may draw two segments for a cross and three segments for an arrow. This opens the way for future studies on gesture recognition based on the coordinates of segment points on the plane and on the association of semantic meanings to gestures (i.e. “crossing out” an entity or using an arrow to point at an entity). We believe that our positive experience using these modalities within NESPOLE! deserves the attention of other developers of multi-modal interfaces.

3.4 The NESPOLE! Monitor

We have found it to be extremely important and useful to provide the users with the ability to monitor the recognition, analysis and synthesis implemented by the translation components of the system in order to keep track of the translation process. The NESPOLE! Monitor has been developed to provide this feedback to the users. The monitor display for each user includes the following fields:

- ‘Remote Speech Translation’: the textual representation of the last translated utterance that arrived from the other party.
- ‘System Hears’: the recognized text representation of the last utterance spoken by the local user, as recognized by the speech recognizer within the HLT server for the language of the local speaker.
- ‘System Understands’: a textual representation resulting from the translation of the last utterance

spoken by the local user back into their own language. The purpose of this field is to provide the user with the ability to identify cases where the translation is likely incorrect, due to incorrect analysis of the spoken input utterance by the translation system.

By monitoring the 'System Hears' field, the user can verify the accurate recognition of the last spoken utterance. Similarly, by monitoring the 'System Understands' field, the user can verify that the meaning of the utterance was correctly captured by the analyzer within the translation server (by judging whether the paraphrase back into their own language reflects the same meaning as the originally spoken utterance). When a translation failure is detected, the user can click on a 'Cancel Translation' button, which generates a red, flashing message on the monitor of the other party, alerting them to the fact that the incoming translated message should be ignored. The user can then repeat or rephrase the message. If multiple recognition attempts of the same sentence fail, the user can manually edit the "System Hears" field, correct the sentence and resend it to the translation server, in order to eliminate the mistake made by the recognizer.

This kind of feed-back has been demonstrated to be very helpful for "expert" users, who are familiar with machine translation technology and have gained some experience with the NESPOLE! system. Recent experiments and users studies [3, 4], however, have demonstrated that even novice users find this type of information very useful and can learn to use the functionalities after a brief training or usage of the system. Some further improvements to the feedback on the state of the system are currently being implemented (see § 5.2).

4. Speech and Gestures Synchronization

Event synchronization is extremely important in multimodal real-time applications. In the NESPOLE! system, the transmission of audio, video and data over the IP network and the delay imposed by translation required the implementation of a synchronization mechanism between the gestures (graphic signs drawn on AeWhiteboard) and the speech occurring while performing the gesture. Lack of synchronization could result in awkward situations in which the speech corresponding to a gesture is heard by the remote user after a significant delay, possibly disrupting the flow of the communication between the two parties. A gesture buffering algorithm within the Mediator was designed to avoid such situations by sending the drawing and the related sentence to the remote user at the same time. The buffering is based on the timestamps that are contained within the audio packets received by the Mediator during the H.323 connection. When the Mediator receives a data packet with gesture information from one of the two systems, it buffers and associates it to the timestamp contained in the received audio packet, rather than sending it to the other system immediately. The audio packet is sent to the Global NESPOLE! Server for translation. When the translation arrives, the Mediator sends it to the target system together with the buffered data. In this way audio and gestures are simultaneously received by the target system.

Buffering within the Mediator is optional, and controllable as a runtime feature. Our experience has shown that the "no buffering" option is more suitable and natural in a "push to talk" mode (see § 5.1), in which gestures are made after (or sometimes before) but not during the speech utterance and therefore there is no need for synchronization.

5. User Studies and Interface Improvement

Usability assessment with novice users plays a basic role in development of well-designed interfaces, with the goal of developing an interface that provides real users an effective and pleasant interaction experience. Throughout the first two years of the NESPOLE! project we took advantage of many opportunities to collect data concerning the system (and the interface) usability with actual users - both computer experts and people with little to no computer skills. Most of the interface improvements leading to the current version were based on the comments and suggestions of these users, and on our observation of their behavior. This process is still in progress as recently collected data has suggested some additional functionalities, which will be implemented in the near future.

5.1 System Evolution and Current Status

Many interface improvements resulted from lessons learned from observations of the experimental interactions (Summer 2001, see [3, 4]) and from the User Group Meeting (September 2001). The goal of the experiment was to evaluate the added value of multi-modal input in the multilingual NESPOLE! system. To run this experiment, we needed to have a steady and usable system, which could support a real "free" interaction among novice users. In support of this goal, we started collecting full monolingual and multilingual dialogues with actual users a few months prior to the experiment. In both cases (monolingual and multilingual dialogues), a researcher assisted the users, observed their behavior and took note of emerging problems and the users' complaints. On the basis of the observations made, we incrementally improved the user interface. The main changes concerned addition of zoom and scroll functions, the possibility of saving a

map with the gestures performed on it, window size optimization and improvement of the users comprehension of each element on the screen.

The experiment conducted in Summer 2001 involved 35 novice users speaking Italian, German or English. The system worked in a "push-to-talk" mode, where the user had to push a button to enable speech input (technically disabling the "mute" function) and then had to push it again at the end of the utterance (to turn on "mute" again). We used the "no-buffering" version of the Mediator, with gestures immediately sent to the remote partner. We compared the multi-modal version of the NESPOLE! system with a system version in which multi-modal resources (pen-based gestures) were not available. We found several indictors suggesting that the multi-modal (MM) version was more effective than the speech-only (SO) one in supporting the interaction. The first indicator is fewer repeated turns in the MM dialogues, especially when spatial information is conveyed (English-Italian MM dialogues contained 11% repeated turn versus 17% for SO; for German-Italian dialogues repeated turns amounted to 18% for MM versus 23% for SO). The second indicator is fewer unsuccessful turns in MM dialogues (19% for MM versus 30% for SO in the English-Italian dialogues; 18% versus 31% in the German-Italian dialogues). The third indicator is the lower number of dialogue segments containing identifiable misunderstanding between the two parties (one such segment in each of 3 of the MM dialogues, versus a total of 7 such segments in the SO dialogues). We concluded therefore that multimodality helped users to overcome ambiguities and system recognition and translation errors **.**

One of the most frustrating experiences encountered during the experiment were time delays due to the very long map transfer times (from about half a minute to about two minutes, depending on the bandwidth and on the network conditions). Even previously shared maps were re-transmitted whenever accessed again later in the communication. This brought us to implement a new mechanism of map transfer: the bitmap file is now actually transferred only if not previously transferred; otherwise, only the name of the bitmap is transferred and the remote system loads it locally. Therefore, even though the experiment was not designed to evaluate the system interface, the interface benefited from the experiment.

The NESPOLE! User Group Meeting involved 8 European and US companies that met in Trento in September 2001. After a demonstration of the system, a questionnaire was distributed to all participants and some discussion followed. The questionnaire consisted in 14 statements which the subject was to give his/her attitude towards by using a five-grades scale plus four open-ended questions. The questionnaire has not been validated yet, and the sample was not wide enough (n=19) to perform statistics, nor we had any intention to provide for such information. Our purpose was simply to receive an overall qualitative assessment of our system's features from potential commercial users. All users agreed that the visualization of maps and the ability to interact directly using gestures significantly enhance the communication, thus confirming the results of the experiment with non expert users (see above). Graphical features were judged reasonably good: all the participants agreed that the functionality of the elements on the screen is clear and that the windows do not create confusion. Nevertheless, 55% of them reported that the elements on the screen are not very pleasant. Questionnaire answers also confirmed our impressions concerning problematic system functionalities: 89% of the participants reported that the system is slow and that the map loading time is too long. Answers given to the open-ended questions and the comments collected during the discussion indicated that the feed-back concerning "what is happening" when original audio is disabled is too poor. We already have plans to improve the quality of this feed-back (see § 5.2 for details).

Other comments received during the User Group meeting resulted in interface improvements. One useful suggestion was to allow the field containing the speech recognizer output to be edited manually. In the early stages of the project, the NESPOLE! Monitor had no editable fields, and users could not correct errors manually. This new added feature has proven to be quite effective in recent system demonstrations, including the IST 2001 Event -"Technologies serving people" in Düsseldorf.

Another observation was that the "push-to-talk" button positioned within the Netmeeting® window, was perceived to be too small and quite difficult to manage. Thanks to the NetMeeting® SDK we managed to implement a larger "push-to-talk" button within the NESPOLE! Monitor window, which is always available to the user during the interaction.

5.2. Current and Future Plans

Further improvements of the NESPOLE! user interface are planned in the near future. The first change (often requested by users) is the addition of live video transmission. Video clearly conveys significant non-verbal information. The availability of video within our speech-to-speech translation system has the potential for improving the quality of the communication by providing natural feedback.

Another planned modification relates to the feedback provided by the NESPOLE! Monitor window. The objective is to differentiate between an expert user and a novice one: we plan to transfer the

contents of the NESPOLE! Monitor editable fields to a Dialog History Window, where they will be enriched with other data arriving from the Global NESPOLE! Server and all text strings will be ordered on a temporal base. The Dialog History Window will be optional and used primarily for system demonstrations. Common users will only see a window containing the output of speech recognizer, which they can edit or cancel (by pressing a "Cancel Translation" button) and two icons indicating system's status.

Some other changes to the interface are planned, based on the questionnaire answers we collected during a monolingual data collection carried out in February 2002. Like the previous, this questionnaire has not been validated yet. The purpose was to collect impressions and suggestions concerning the most relevant system features: instructions, screen and devices, multimedia and multimodality, usefulness and effectiveness of the system. The questionnaire consisted of 16 statements, to which the subject was to give his/her attitude towards by using a four-graded scale (from complete agreement to complete disagreement) and four free-reply items.

We collected 77 questionnaires (65 from volunteers who took part in one dialogue playing the role of the customers and 12 from APT tourist office agents). The overall judgement of the system was good. 100% of the items received an overall positive judgment (strong or mild agreement on positive items and strong or mild disagreement on negative items); 100% of the users judged maps and drawings a very helpful support to the communication; 91% of the volunteers acting as customers reported that they would like to use the NESPOLE! system instead of telephone or e-mail to ask for tourist information, if it were available on the web; 100% of the tourist agents reported they would like to use the system at work.

As to usability of single features, the main relevant issue is the map saving procedure: 8 agents (75%) mentioned that the map saving procedure is too slow and that it interrupts the dialogue flow: the user saving the map has to choose a file name and to select a directory where the map has to be saved. We therefore plan to implement a feature that will allow the map files to be saved by simply clicking a button.

6. Conclusions

The NESPOLE! user interface described in this paper supports a friendly, pleasant, useful and interactive system that can be used by a variety of users with very different levels of instruction and computer skills. Our system allows users speaking different languages to interact in a natural way, by means of:

- hearing their original voice and their translated speech;
- seeing the face and expression of their partner;
- exchanging data information through gestures on drawings and map selections;
- monitoring the translation process and correcting it if and when something goes wrong.

What is very innovative about our system is the means by which the user can perform these functions: a light PC with standard free software for video-conferencing (Microsoft® NetMeeting®), a microphone and a camera: common users should be able to use the NESPOLE! Service from their home PC.

We are planning to install and test the NESPOLE! system at the APT of Trento, by the end of 2002. Initially only a monolingual version of the system will be made available. We will then consider enabling the multilingual version of the system for general use.

7. Acknowledgements

This work was supported by NSF Grant 9982227 and EU Grant IST 1999-11562 as part of the joint EU/NSF MLIAM research initiative.

We would thank the NESPOLE! people who gave a contribution to the Interface design and the data collection, and in particular Francesca Guerzoni, who developed the first version of the usability questionnaire and analyzed the data collected during the User Group Meeting.

8. References

[1] A. Lavie, F. Pianesi, et al. "Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications". In *Proc. of the HLT2001*, San Diego, CA, 2001. ACM.

[2] F. Metze, C. Langley, A. Lavie, J. McDonough, H.Soltau, L.Levin, T.Schultz, A. Waibel, R. Cattoni, G. Lazzari, N.Mana, F. Pianesi, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei. "The NESPOLE! Speech-to-Speech Translation System". In *Proceedings of Human Language Technology Conference (HLT-2002)*, 3 2002.

[3] E. Costantini, F. Pianesi, S. Burger. "The Added Value of Multimodality in the NESPOLE! Speech-to-Speech Translation System: an Experimental Study". In proceedings of ICMI 2002.

[4] *NESPOLE! Project Deliverable D8 – First Showcase Documentation.* Available on NESPOLE! Project website: http://nespole.itc.it