The NESPOLE! System for Multilingual Speech Communication over the Internet

Alon Lavie, Fabio Pianesi and Lori Levin

***Abstract*— The NESPOLE! System is a speech communication system designed to support multi-lingual interaction between common users and providers of e-commerce services over the internet. The core of the system is a distributed interlingua-based speech-to-speech translation system, which is supported by multi-modal capabilities that allow the two parties participating in the communication to share web pages and graphical content which can be annotated using gestures. We describe the unique features and considerations behind the design and implementation of this system, and evaluate these within the context of a constructed full prototype of the system that was developed for the domain of travel planning.**

***Index Terms*— Distributed Processing, Machine Translation, Multimodal Interfaces, Speech Communication.**

I. Introduction

THE NESPOLE! System is a speech communication system designed to support multilingual interactions between common users with standard personal computers and providers of e-commerce services over the internet. The core of the system is a distributed interlingua-based speech-to-speech translation architecture, which is supported by multimodal interfaces. The multimodal capabilities allow the two parties participating in the communication to share web pages and graphical content, such as maps, diagrams and pictures. Both parties can annotate the graphics using gestures, which are visible simultaneously on both sides. The multimodal capabilities result in significantly improved communication robustness.

The NESPOLE! System was developed in the course of a large collaborative project funded by the European Commission and the US National Science Foundation (NSF). The project involved three European research labs (ITC-irst in Trento Italy, ISL at University of Karlsruhe in Germany, CLIPS at Université Joseph Fourier in Grenoble France), one research group in the US (ISL at Carnegie Mellon University), and two industrial partners (APT - the Trentino provincial tourism bureau, and AETHRA - an Italian tele-communications company). The NESPOLE! system supports communication between Italian-speaking human agents and clients speaking English, German or French, in the limited albeit large domains of travel planning and medical assistance. While similar in its domain to prototype systems developed in C-STAR [1] and VERBMOBIL [2], the NESPOLE! system is groundbreaking in several important aspects:

- *Distributed Architecture:* Extremely little hardware and software reside on the PCs of the parties engaged in communication. The communication channel between the two parties is separated from the speech translation sub-system. The speech translation components of the system are distributed into language specific servers, which can be developed and maintained independently.
- *Speech Recognition and Translation over the Internet:* Audio signal quality is affected by being captured at the PCs of end users, shipped over the internet and recognized at remote servers. The speech translation components are designed to be robust to the resulting degraded output and are distributed over the internet, allowing them to be physically located in locations that are advantageous for performance.
- *Interlingua Design:* We designed a new interlingua representation for the project that is specifically suitable for limited task-oriented domains. This interlingua is sufficiently rich to capture speaker intention, simple enough to be used reliably by developers of different languages working independently, and flexible enough to support advanced techniques for analysis and generation.
- *New Approaches to Language Analysis into the Interlingua Representation:* For English, German and Italian, new approaches to language analysis were developed, using machine-learning and classification methods that significantly reduce manual development time and improve domain portability.
- *Integration of Speech Translation with Multi-modal Communication:* The architecture of the system embeds

Manuscript received June 21, 2004. This work was supported in part by the U.S. National Science Foundation under Grant IIS-9982227 and the European Commission under Grant IST-1999-11562 as part of the joint EU/NSF MLIAM research initiative.

Alon Lavie is with the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, PA 15213 USA (corresponding author: phone: 412-268-5655; fax: 412-268-6298; e-mail: alavie@cs.cmu.edu).

Fabio Pianesi is with Instituto Trentino di Cultura – Centro per la Ricerca Scientifica e Tecnologica (ITC-irst), Trento, 38100, Italy (email: pianesi@itc.it).

Lori Levin is with the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, PA 15213 USA (email: lsl@cs.cmu.edu).

the translated speech communication between the parties within multi-modal interfaces that significantly enhance communication robustness.

In this article, we describe the main principles behind the design and implementation of the above aspects of the NESPOLE! System, and discuss these within the context of a prototype of the system in the domain of travel planning that was constructed in the course of the project. A second prototype of the system was constructed in the Medical Assistance domain in order to explore the domain portability of the system's architecture and components. The design and evaluation of these portability experiments fall beyond the scope of this article.

The remainder of this article is organized as follows. Section II provides an overview of the architecture of the system, including the distributed speech translation modules and the integration of machine translation with the speech and multi-modal communication channel between the two parties involved in the communication. Section III presents an overview of the multimodal interfaces in the system. Section IV describes the design of the Interlingua representation and how this design supports the requirements of our system. Section V provides an overview of one approach for analysis of source language input into interlingua representations. Section VI presents a comprehensive evaluation of the system from several different perspectives, including the robustness of the system to network traffic conditions, a detailed assessment of the effectiveness of multi-modal communication, and an evaluation of end-to-end translation quality at the sentence-level. Finally, we present our conclusions in Section VII.

II. System Design and Architecture

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. Standard commercially available PC video-conferencing technology, such as Microsoft's NetMeeting®, is used to connect between the two parties in real-time. The communication is initiated by the client clicking on a dedicated "button" within the web-page in order to establish the video-conferencing connection with the agent of the provider site. The client is then presented with an interface consisting primarily of a standard video-conferencing application window and a shared whiteboard application. Using this interface, the client can carry on a conversation with the agent, where the NESPOLE! system provides two-way speech-to-speech translation between the parties. The agent speaks Italian, while the client can speak English, French or German. A short sample dialogue between an English client and an Italian agent is shown in Figure 1.

A: Buongiorno
(Good morning)
C: Good morning. I would like to visit Val di Fiemme.
A: Le faccio vedere una mappa della Val di Fiemme.
(I'll show you a map of Val di Fiemme.)
[agent shows map of Val di Fiemme]
A: Riesce a vederla?
(Can you see it?)
C: Yes, I can see the map.
A: Aveva in mente un posto in particolare?
(Do you have a particular location in mind?)
C: No.
A: Ci sarebbe un pacchetto per una settimana, per due persone, in un hotel a tre stelle.
(There is a one week package for two people in a three star hotel.)
A: L'hotel si trova a Panchià.
(The hotel is located in Panchia.)
[agent indicates Panchià with a red square]
C: Could I go cross-country skiing?
A: Certamente. Le mostro le piste da fondo sulla mappa con un cerchio blu.
(Yes. I am showing you the cross country skiing on the map with a blue circle.)
[agent indicates the slopes with a blue circle]
Riesce a vederlo?
(Can you see it?)
C: Yes, thanks.
A: Desidera qualcos'altro?
(Is there anything else?)
C: No, thanks.
A: Grazie a lei. Arrivederci.
(Thank you. Goodbye.)

Figure 1: Sample English/Italian Dialogue in the Travel Planning Domain

A schematic diagram of the NESPOLE! system architecture is shown in Figure 2. 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. The HLT servers provide the actual speech recognition and translation capabilities. This system design allows for a flexible and distributed architecture: mediators and HLT-servers can be run in various physical locations, so that the optimal configuration, given the locations of the client and the agent and anticipated network traffic, can be taken into account at any time. An API allows the HLT servers to communicate with each other and with the mediator, while the HLT modules within the servers for the different languages are implemented using very different approaches and software. Further details of the design principles of the system are described in [1].

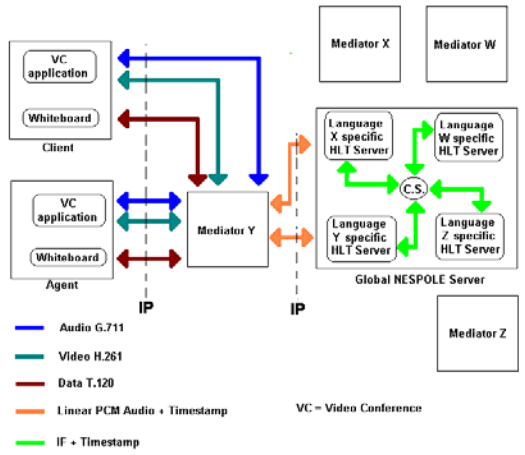

Figure 2: General Architecture of the NESPOLE! System

The system architecture shown in Figure 2 contains two different types of Internet connections with different characteristics. The connection between Client/Agent PCs and the Mediator is a standard video-conferencing connection that uses H323 and UDP protocols. In cases of insufficient network bandwidth, these protocols compromise performance by allowing delayed or lost packets of data to be "dropped" on the receiving side, in order to minimize delays and ensure close to real-time performance. The connection between the Mediator and the HLT servers uses TCP over IP in order to achieve lossless communication between the Mediator and the translation components. For practical reasons, Mediator and HLT servers in our system usually run in separate and distant locations, which can introduce some additional time delay. System response times in recent demonstrations have been about three times real-time.

The global NESPOLE! HLT server comprises four separate language-specific servers. Additional language-specific HLT servers could easily be integrated. 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.

III. Multi-modal Interfaces and Capabilities

The NESPOLE! user interfaces are displayed at both client and agent PCs. The interface is composed of 4 windows: a web browser, a Microsoft® NetMeeting® video-conferencing interface, the AeWhiteboard and the NESPOLE! Monitor. Each window plays a different role in the interaction between client and agent and their activation is automatic. We present an overview of the functionality of each component below. Further details can be found in [4].

A. NetMeeting

Once the system is activated, the user can hear and see the agent "live" in the standard Netmeeting® interface. The two users can then communicate, each speaking their own language. 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.

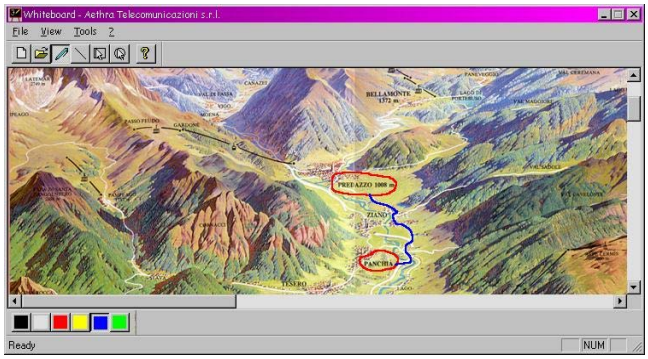

Figure 3: The NESPOLE! AeWhiteboard Application

B. The AeWhiteboard

The AeWhiteboard (Figure 3) 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 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.

The AeWhiteboard drawing functionalities include free-hand strokes, where the user can draw arrows, lines, and circles; lines that can connect two points on the map; and highlighted areas, where the user can enclose areas on the map with within an elliptical/rectangular box. The drawings are performed by means of a tablet-pen device or a standard mouse. Different gestures can utilize different colors and shapes. The user can save a copy of the map along with any annotations for later reuse. Another important function 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.

The notion of collaborative shared workspaces and interfaces has been long pursued by the Computer Supported Cooperative Work (CSCW) community and has become pervasive technology in both business and e-learning [5]. Results from experiments [6], also reported later in this article, 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.

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.

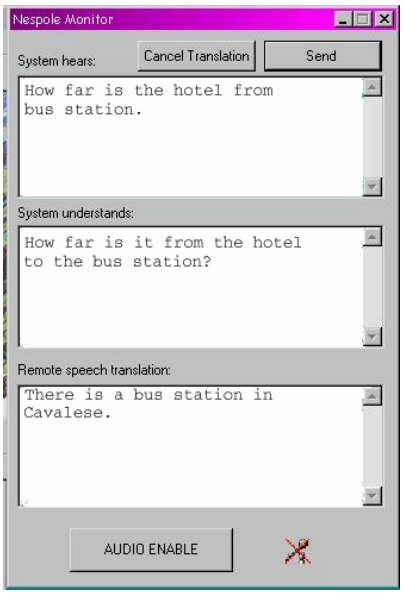

Figure 4: The NESPOLE! Translation Monitor

C. The NESPOLE! Monitor

We have found it 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 (Figure 4) was 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.

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. Our experiments and users studies [6] indicate that even novice users find this type of information useful and can learn to use the provided functions after a brief training or usage of the system.

IV. THE NESPOLE! INTERLINGUA REPRESENTATION

An interlingua is a representation of meaning or communicative intention. An ideal interlingua is neutral between the different syntactic expressions of the same meaning in different languages. Designing such an interlingua is notoriously difficult, and this has been the focus of significant research over the past two decades [7] [8] [9] [10]. One advantage of an interlingua approach is that the analyzers and generators can be written by mono-lingual system developers. Interlingua MT is particularly advantageous when more than two languages are involved. To add a new language, an analyzer and generator are simply connected to the interlingua. Interlingua-based MT also supports paraphrase of the input in the original language. This is extremely useful in speech translation, where language analysis is much more error prone than generation. Once the analyzer converts an input sentence into its appropriate interlingua, the interlingua can be generated back into the language of the original speaker and the speaker can then verify that the meaning captured in the analysis process is in fact reasonably correct.

The interlingua used in the NESPOLE! system is called Interchange Format (IF) [11], [12], [13]. The IF defines a shallow semantic representation for task-oriented utterances

that abstracts away from language-specific syntax while capturing the meaning of the input. Each utterance is divided into semantic segments and an IF is assigned to each segment. An IF representation consists of four parts: a *speaker tag*, a *speech act*, an optional *sequence of concepts*, and an optional *set of arguments*. The representation takes the following form:

speaker : speech act +concept* (argument*)

The speaker tag indicates the role of the speaker in the dialogue. The speech act captures the speaker's intention. The concept sequence, which may contain zero or more concepts, captures the focus of a semantic segment. The speech act and concept sequence are collectively referred to as the *domain action* (DA). A specification document developed in the course of the project defines all of the components and describes how they are allowed to combine into valid IF representations. Several examples of utterances with corresponding IFs are shown in Figure 5.

Thank you very much.

```
a:thank
```

Hello.

```
c:greeting (greeting=hello)
```

How far in advance do I need to book a room for the Al-Cervo Hotel?

```
c:request-suggestion+reservation+room
  (suggest-strength=strong,
   time=(time-relation=before,
         time-distance=question),
   who=i,
   room-spec=(room,
      identifiability=no,
       location=
        (object-name=cervo_hotel)))
```

Figure 5: Examples of Utterances and their Corresponding NESPOLE! Interlingua Representations

The NESPOLE! Interlingua is based on representing the speaker's intention rather than the literal meaning of the utterance. This makes the IF particularly suited to resolving translation mismatches that are the result of a meaning being expressed using different syntactic structures in different languages [7]. Representing speaker intention instead of literal meaning (or predicate argument structure) also contributes to the language-neutrality of the IF. Many speech acts are expressed by formulaic utterances that do not translate literally into other languages. The design of the NESPOLE! IF attempted to carefully balance several requirements that at times compete with each other. These include the range of meanings that can be expressed, portability to other semantic domains and inter annotator agreement.

A significant number of dialogues were manually tagged with IF representations, and this data was later used for training language analysis components. The NESPOLE! IF database [14] contains around 5000 sentences in English, Italian, and German that are tagged with IF. Although the IF is specific to the NESPOLE! project, the database is a valuable resource that can be used as a research tool in many areas of language technologies.

V. Data-Trainable Language Analysis

One advantage of the distributed interlingua-based design of the NESPOLE! System is that it allows independent and radically different approaches to language analysis and generation for the various languages. In this section, we present an overview of only one new approach to language analysis, which was developed for English and German analysis in our system. The analysis approach uses a hybrid combination of grammar-based parsing and machine learning techniques to transform spoken utterances into the IF representation described above. The speaker tag is assumed to be given. Thus, the goal of the analyzer is to identify the DA and arguments. The hybrid analyzer performs argument parsing by using a robust parser and handwritten phrase-level semantic grammars to extract low-level interlingua arguments from input utterances. The phrase level grammars used by the analyzer are easier to develop and less domain-dependent than semantic grammars that parse full domain actions. The analyzer then uses automatic classification techniques to divide an utterance into semantic segments and to determine the domain action for each semantic segment. This domain action classification task can be performed with reasonably high accuracy and improves the robustness of the analyzer to unseen in-domain inputs and inputs from an automatic speech recognizer. Furthermore, applying machine learning techniques to the task of domain action classification improves portability by eliminating the most labor-intensive stage of grammar writing and reducing data annotation requirements. Similar hybrid techniques involving shallow parsing and classification of natural language input have been used in other language processing tasks such as Call Routing [27] [28]. The Italian analysis module used in the NESPOLE! system uses a similar approach, where trained language models are used for DA classification [29].

The first stage in analysis involves parsing an utterance for arguments (specific times, locations, room-types, etc.). During this stage, utterances are parsed with phrase-level semantic grammars using the robust SOUP parser [18].

The second stage of processing in our hybrid analysis approach is segmentation of input utterances into semantic units. Recognized utterances may consist of several semantic segments, each of which must be assigned an IF. Utterances must therefore be split into segments before the domain actions can be determined. Since segmentation is performed on the output from a speech recognizer, neither punctuation nor case information is available during segmentation, and some input utterances may contain speech recognition errors.

The semantic segmenter makes a binary decision about the presence or absence of a boundary at each potential boundary position in the argument parse output. We use TiMBL [19], a

memory based (k-Nearest-Neighbor) for the segmentation classifier. The input to the TiMBL segmentation classifier consists of a set of 10 features based on the word and argument parse information surrounding a potential boundary position. The output of the classifier is a binary decision about the presence or absence of a segment boundary at the position. Classifiers were trained on about 7000 utterances for English and about 15000 utterances for German, and tested using a "leave-one-out" strategy. For both languages, the Classifier achieves about 93% correct boundary classification.

Following argument parsing and semantic dialogue unit segmentation, the third stage of processing in the hybrid analysis approach is domain action classification. The purpose of this stage is to identify the domain action for each semantic dialogue unit in an input utterance. The output of the classifier is checked against the IF specification to ensure that a legal IF representation is produced for each segment.

One classifier is used to identify the speech act, and a second classifier identifies the concept sequence. Both classifiers are implemented using TiMBL [19], a memory-based learner. Speech act classification is performed first. Input to the speech act classifier is a set of binary features that indicate whether each of the possible argument and pseudo-argument labels is present in the argument parse for the segment. No other features are currently used. Concept sequence classification is performed after speech act classification. The concept sequence classifier uses the same feature set as the speech act classifier with one additional feature: the speech act assigned by the speech act classifier. The classifiers are trained on data extracted from the NESPOLE! database of IF-tagged utterances. Table I contains the characteristics of the training data for the travel domain. Classification accuracies for speech acts, concept sequences and full DAs were calculated by a 20-fold cross-validation experiment, and are shown in Table II. Complete details of the analysis approach described above can be found in [20] [21] [22].

Table I: Characteristics of Travel Domain Training Data for Domain Action Classification

	English	**German**
Semantic Segments	8289	8719
Domain Actions	972	1001
Speech Acts	70	70
Concept Sequences	615	638
Vocabulary	1946	2815

Table II: Travel Domain Classification Accuracy Levels for Speech Acts, Concept Sequences and Domain Actions

	English	**German**
Speech Acts	79.00	77.16
Concept Sequences	67.97	67.08
Domain Actions	56.82	54.96

VI. Performance Assessment and Evaluation

We evaluated the performance of the NESPOLE! system along several different dimensions. We summarize here results for the following types of assessment:

1. The impact of network traffic and the consequences of real packet-loss on system performance.
2. The impact of multi-modal communication.
3. End-to-end translation performance evaluations at the level of semantic dialogue units.

A. Network Traffic Impact

In our various user studies and demonstrations, we were forced to deal with the detrimental effects of network congestion on the transmission of Voice-over-IP in our system. The critical network paths are the H323 connections between the Mediator and the client and agent PCs. These connections use the UDP protocol in order to guarantee real-time human-to-human communication, but may consequently loose packets of information. This can potentially be very detrimental to the performance of speech recognizers [23]. The communication between the Mediator and HLT servers uses TCP, which can introduce time delays, but no packet loss.

To quantify the influence of UDP packet-loss on system performance, we ran a number of tests with German client installations in the USA and Germany calling a mediator in Italy, which in turn contacted the German HLT server located in Germany. These tests were conducted at different times of the day on different days of the week, in an attempt to investigate a wide variety of real-life network conditions. Sixteen tests were conducted. In each test, a high-quality recording of several hundred development data sentences was channeled into the system on one end (simulating a real user) and processed through the system.

Results indicated that our speech recognition engine is relatively robust to packet loss rates of up to 5%. There was no clear degradation in the word accuracy of the recognizer as a function of packet loss rate in this range. Word accuracy levels ranged between 63% and 59% for packet-loss conditions between 0.1% and 5.2% (the word accuracy on clean 16kHz recording is 71.2%). Our experience indicates that packet loss rates of over 5% are quite rare under normal network traffic conditions. For detailed results, see [24].

B. Evaluation of Multi-modal Communication Effectiveness

We conducted three user studies aimed at comprehensively assessing the communication effectiveness of the multi-modal capabilities of the developed system. The first study compared multilingual dialogues with and without the ability to perform multimodal gestures [6]. The second study compared multilingual with monolingual dialogues, and the effects of "push-to-talk" restrictions on dialogue structure [25], [26]. The third study investigated overall system usability issues, based on questionnaires filled out by users after using the

system. We focus here on the results of studies 1 and 2 and highlight conclusions from study 3.

1) Study-1: Effects of Multi-modal Interactions

The first experiment was designed to test whether the availability of gestures increased the probability of successful interaction, by a) reducing the frequency of miscommunication and disfluency; and b) supporting a faster recovery from recognition and translation errors. Two experimental conditions were devised: a speech-only condition (SO), involving multilingual communication and the sharing of images; and a multimodal condition (MM), where users could additionally convey spatial information by means of pen-based gestures on the shared maps. We collected 28 dialogues, fourteen featuring an English customer, and fourteen a German one; in all cases the agent was Italian. Each group was further divided into seven SO dialogues and seven MM ones. The customer's task was to choose an appropriate location and a hotel within specified constraints concerning the relevant location, the available budget, etc. The agent's task was to provide the necessary information. We refer the reader to [6] for more details about this study; here we report only the results relating to task completion and success rate, focusing on the English-to-Italian data.

Both the SO and MM versions of the system were equally effective for task completion: in both cases, 86% of the users were able to complete the task. Hence, the NESPOLE! system was sufficient for novice users to accomplish a task with minimal written instructions, very short initial training on the interface, and no further assistance during the interaction.

SO and MM dialogues differed in terms of unsuccessful and repeated turns, particularly so in the spatial segments of the dialogues. For instance, when spacial information was involved, MM had a lower failure rate than SO (19% vs. 30%), and fewer repeated turns (11% vs. 17%). At a global level, MM had a higher return rate than SO (31 vs. 19). Here "returns" refer to cases in which previously addressed, but non-completed, topics are reconsidered later on. The return rate is, therefore, a measure of dialogue fluency. Finally, only 21% dialogues in MM versus 50% in SO contained misunderstandings related to, or originating from place names. Moreover, misunderstandings arising in MM were most often immediately resolved using gestures, while in SO ambiguous or misunderstood sub-dialogues often remained unresolved.

2) Study-2: Effects of Translation and Push-to-talk on Communication and Dialogue Structure

The goal of the second study [25] [26] was to investigate in greater detail the impact of the NESPOLE! system, with all its delays, translation errors and technical problems, on the dialogue structure produced by subjects and on the way gestures are integrated with speech. We wanted also to isolate and quantify the effects of the push-to-talk restriction on dialogue structure and flow within our system.

We consequently designed three experimental conditions: (1) *the STST multilingual condition*, in which the users communicated by means of the NESPOLE! system in a push-to-talk mode and with speech translation (English/Italian) enabled; (2) *the PTT condition*, in which the system was used, but for monolingual (Italian/Italian) communication, again in a push to talk mode; and (3) *the non-PTT condition*, identical to (2), but without push-to-talk.

Seven English/Italian multilingual dialogues, eight Italian PTT dialogues and eight Italian non-PTT dialogues were collected and analyzed. For the STST condition, seven English speaking customers located in Pittsburgh interacted through the NESPOLE! system with three tourist agents located in Italy. For the PTT and Non-PTT dialogues both the agents and the clients were located in Italy, resulting in better network connections and very limited transfer delays.

Transcriptions of the data included annotations for spontaneous phenomena, dialogue structure, and gestures. For the dialogue structure, we used a modified version of the Dialogue Structure Coding Scheme (DSCS) [27] [28], adapting it to the needs of our scenario, while keeping to the idea that dialogues are to be analyzed in terms of games, the latter in turn being formed by moves. We refer the reader to [25] for a more complete description of the modified DSCS. Additional levels of annotation were used to kept track of whether a move was continued, abandoned, repeated, or reformulated, and whether the move concerned technical issues (e.g. bad audio) or multimodal issues.

The collected corpus consisted of a total number of 18,100 word tokens. The average duration of a dialogue was 23 minutes for STST, 9.85 minutes for PTT, and 8.87 minutes for Non-PTT. The difference in dialogue duration between the monolingual conditions and the multilingual one is attributable to: (1) the time delays introduced by for the speech translation process in our system, and (2) the Internet's rate of information transfer. Silence, translation and speech synthesis accounted for 87% of the dialogue duration in STST, 49% in PTT and 19% in Non-PTT. We performed a multivariate analysis of variance on talk time, number and types of tokens produced, number of turns and frequency of spontaneous phenomena, with condition and user (agent vs. client) as factors. Agents spoke longer, with more turns, tokens and token types than clients (this reflects the nature of the task), and produced more spontaneous phenomena (coughs, pauses, breaths, etc.). As to the three conditions, we found the ordering STST < PTT < Non-PTT on all variables except for number of turns, where the differences between STST and PTT were not statistically significant; both were lower than Non-PTT. Hence, our subjects spoke longer, with more tokens and token types occurring when the communication set-up was more "natural" (exemplified by the Non-PTT condition). In STST subjects spoke in a more controlled and planned way, and this was reflected by the amount of spontaneous phenomena. The differences detected in all the above variables were statistically significant ($p<.05$).

An analysis of the discourse structure in terms of games and

moves across the three conditions indicated that games tended to be shorter in STST and PTT and longer in Non-PTT. This can be interpreted as indicative that the push-to-talk mode has a significant effect on the way dialogues are structured at the global level. On the other hand, the analysis showed that there was a trend towards fewer nested games (games embedded within another game) in the STST condition (10% of the games) than in the two monolingual conditions (26% in PTT and 23% in the Non-PTT condition). This indicates a more complex dialogue structure in the monolingual conditions. It appears that while the push-to-talk constraint has a major effect on the global structure of the dialogue, the translation system affects the internal structure, resulting in simpler internal structures in STST (fewer nested games).

Moves with similar functions were grouped together in broader categories. The resulting classes were "query", "information", "check/align" (which includes two moves whose function is to check for comprehension and transfer success), "acknowledgement" (acceptance), "action" (the description of an action or gesture), "ready" (preparation), and "other" (which consists of less frequently occurring moves).

We ran a logistic regression analysis of the data (overall significance level $p<.05$, with correction for multiple comparisons) with the number of moves as the dependent variable, and condition and user (agent and client) as the independent variables. Here we focus on the effects of condition. Both STST and PTT had a lower number of acknowledge moves than Non-PTT. Acknowledges were used as game-closing moves 66% of the times in Non-PTT, while the figures for PTT and STST are 38% and 23%, respectively. Query moves increased in STST with respect to Non-PTT; the PTT vs. STST and PTT vs. Non-PTT comparisons did not yield statistically significant results. The percentage of turns that provide information are similar (around 30%) in all conditions. Noticeably, STST is the only condition having approximately the same number of information-requesting and information-providing-moves; in the monolingual conditions the frequencies of the former are lower than those of the latter. This suggests that the amount of spontaneously offered, not elicited information is higher in PTT and Non-PTT than in STST, and that in STST a structure prevails where information-providing moves are more often elicited by information requests. Finally, the information moves show an opposite trend to that of acknowledge ones: 25% of the games ends with an information move in Non-PTT, 52% in PTT and 50% in STST condition.

These results can be interpreted as showing that the dialogue structure is simpler in STST than in Non-PTT. People tend to focus on the task (a higher number of simplified question-answer sequences), and to avoid moves, such as acknowledges, that have a more regulatory purpose without directly contributing to the task accomplishment. None of those differences seem to be related to the push-to-talk mode, but seem to reflect a distinction between STST and the two monolingual conditions.

The average number of gestures per dialogue was similar in all three conditions (12.9 in STST, 13.6 in PTT, and 13.7 in Non-PTT); about half were drawings. We distinguished three patterns of temporal integration between gestures and speech: gestures performed (a) immediately before, (b) during, or (c) immediately after the corresponding speech turn. The following table reports the percentages for each category.

Table III: Percentages of Gestures Performed Before, During or After the Corresponding Turn

	STST	**PTT**	**Non-PTT**
Before	32%	8%	0%
During	14%	61%	96%
After	53%	31%	4%

We briefly present now the results of a log-linear analysis (Poisson model) on the relevant data, focusing on statistically significant results (overall level $p<.05$). The number of gestures simultaneous to the speech was lower in STST than in Non-PTT and PTT. In STST about half of them followed the speech (53%), with the content of the turn often anticipating them, e.g., "I'll show you the ice skating rink on the map". In addition, a significant portion of gestures (32%) were performed before the speech. As to PTT the number of gestures performed during speech was significantly lower than those performed in Non-PTT, with a parallel increase of after-speech gestures, which is comparable to the level observed in STST.

Hence, with respect to the condition closer to a natural multimodal dialogue (Non-PTT), the synchrony between gestures and speech was substantially lost in STST. The push-to-talk mode seems to be somehow related to the increase of after-speech gestures. On the other hand, the high number of before-speech gestures in STST seems to be an effect attributable to the translation system itself, and the delays it introduces.

Study-3: System Overall Usability

In a third user study, a post-session questionnaire was given to the participating subjects (nine customers and four agents). The experimental set-up, task, etc., were identical to those of the second study reported above. The questionnaire consisted of twenty questions inquiring about the system's usability and the users' interaction experience. Four items were free-response ones; the other sixteen items required users to mark their level of agreement with respect to a number of statements. Since the questionnaire was not validated, and given the low number of subjects to whom it was administered, we cannot produce objective and valid usability scores. Here we will comment only some qualitative results from the closed-end questions, by distinguishing only between the agreement and disagreement categories.

The system interface was judged positively by both customers and agents. Customers perceived the multi-modal features in NESPOLE! more helpful than agents.

Approximately the same proportion (50%) of customers and agents found the translation quality reasonable. Customers found the system more effective and more helpful than agents. In light of the responses to the open questions, and considering that our agents were all true professional tourism agents, this can be interpreted as indicating that the NESPOLE! system was perceived as more useful and helpful by the people (customers) not professionally involved in multilingual communication.

C. End-to-end Translation Performance Evaluations

Several comprehensive end-to-end translation performance evaluations were conducted in the course of the project. For the travel domain, performance was evaluated at the end of the first year of the project (Showcase-1) and eighteen months later (Showcase-2a). The primary goal of these two evaluations was to assess the improvement in system performance as a function of development effort, as the scale of the domain expanded. In each evaluation, performance was assessed on previously unseen test dialogues. The evaluations were all end-to-end, from input to output, not assessing individual modules or components. We performed both monolingual evaluation (where generated output language was the same as the input language), as well as cross-lingual evaluation. We evaluated on both manually transcribed input as well as on actual speech-recognition of the original audio. Translations were graded by multiple human graders. For each data set, one grader first manually segmented each utterance into semantic segments. All graders then used this segmentation in order to assign scores for each segment in the utterance. Grading is performed using a four-point scale: "very good", "good", "bad" or "very bad". Graders were trained on how to distinguish between the categories and were tested for agreement. We calculate the percent of semantic segments that are graded with each of the above categories. "very good" and "good" percentages are also summed together into a category of "Acceptable" translations. We present here results for the two travel domain systems (Travel-1 and Travel-2). Results shown here are for translation from English and German to Italian, evaluated on client utterances. It is important to note that different test sets were used in the two evaluations, and the results are therefore not fully comparable. In particular, the test set for Travel-2 contained dialogues from a substantially larger set of travel scenarios. Speech recognition word accuracy rates are shown in Table IV. Table V shows the results of translation from transcribed input and Table VI shows the results of translation for Speech-recognized input. Percentages indicate the percentage of translations graded as "acceptable".

Table IV: Speech Recognition Word Accuracies

Language	Travel-1	Travel-2
English	62%	56%
German	64%	51%

Table V: Translation Performance on Transcribed Input

Language	Travel-1	Travel-2
English-to-English	58%	68%
English-to-Italian	55%	70%
German-to-German	46%	61%
German-to-Italian	32%	52%

Table VI: Translation Performance on Speech Input

Language	Travel-1	Travel-2
English-to-English	45%	50%
English-to-Italian	43%	50%
German-to-German	40%	51%
German-to-Italian	27%	53%

While the overall percentage of semantic segments that have an acceptable translation is not extremely high, these numbers do not reflect the dynamic collaborative nature of communication using our system. As reported in the previous section, our usability studies indicate that task completion rates using the system are at levels of around 85%. Complete evaluation results for the various language pairs in the project can be found in [24] and [29].

VII. Conclusions

The NESPOLE! system significantly advanced the state-of-the-art of speech-to-speech translation in several important ways. The architectural design of the system specifically addresses a real-world e-commerce scenario, where speech-to-speech translation is made available as a mode of communication for common users, with standard PCs, no additional hardware and little additional software resident on the client side. The interlingua-based approach and the distributed nature of the system provide flexibility, but impose significant technical challenges that were addressed in effective and innovative ways in the course of the project. Our experiments and user studies indicate that the integration of speech translation with additional modalities of communication significantly enhance the communication robustness of the system. The interlingua design and new approaches to language analysis substantially reduce the development time and effort required to expand the coverage of the system to new domains. Much more work, however, remains to be done in further improving the scalability and portability of the underlying speech translation technology, in order to enable the broad commercial use of such systems in practical, real life situations.

Acknowledgment

The authors wish to thank all members of the NESPOLE! Project at the various participating sites for their dedicated hard work and contributions, without which it would have been impossible to successfully develop the system reported

in this article.

REFERENCES

[1] Levin, L., A. Lavie, M. Woszczyna, and A. Waibel, "The JANUS-III Translation System". *Machine Translation, 15(1-2)*. Pages 3-25.

[2] Wahlster, W. 2000. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin: Springer.

[3] A. Lavie, C. Langley, A. Waibel, , F. Pianesi, G. Lazzari, P. Coletti, L. Taddei and F. Balducci. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Applications. In *Proceedings of Human Language Technology Conference* (*HLT-2001)*, San Diego, CA., 2001. Pages 31-34.

[4] L. Taddei, E. Costantini and A. Lavie. "The NESPOLE! Multimodal Interface for Cross-lingual Communication: Experience and Lessons Learned". In *Proceedings of IEEE International Conference on Multimodal Interfaces (ICMI-2002)*, Pittsburgh, PA, October 2002. Pages 223-228.

[5] Jonathan Grudin, Computer-Supported Cooperative Work: History and Focus, Computer, 27 (5), May 1994. Pages 19-26.

[6] 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 IEEE International Conference on Multimodal Interfaces (ICMI 2002)*. Pittsburgh, PA. October., 2002. Pages 235-240.

[7] B. J. Dorr. *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge, Massachusetts. 1993.

[8] Nirenburg, Sergei, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman. 1992. Machine Translation: A Knowledge-Based Approach. San Mateo, California: Morgan Kaufmann.

[9] Levin, Lori and Sergei Nirenburg. "The Correct Place of Lexical Semantics in Interlingual MT". In *Proceedings of 15th International Conference on Computational Linguistics (COLING-94)*. Kyoto, Japan. Pages 349-355.

[10] Dorr, Bonnie J. "Machine Translation Divergences: A Formal Description and Proposed Solution". *Computational Linguistics, 20(4)*. Pages 597-633.

[11] 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 International Conference on Speech and Language Processing (ICSLP-98)*, Sydney, Australia, November 1998. Volume 4, Pages 1155-1158.

[12] 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 *Proceedings of ANLP/NAACL-2000 Workshop on Applied Interlinguas*, Seattle, WA, April, 2000. Pages 18-23.

[13] L. Levin, D. Gates, D. Wallace, K. Peterson, A. Lavie, F. Pianesi, E. Pianta, R. Cattoni and N. Mana. "Balancing Expressiveness and Simplicity in an Interlingua for Task-based Dialogue". In *Proceedings of Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, Philadelphia, PA, July 2002. Pages 53-60.

[14] L. Levin, D. Gates, D. Wallace, K. Peterson, E. Pianta, and N. Mana. "The NESPOLE! Interchange Format". Deliverable D13. Avialable on NESPOLE! Website at: http://nespole.itc.it

[15] A.L. Gorin, G. Riccardi and J.H. Wright. 1997. "How May I Help You?". Speech *Communication, vol. 2(3)*. Pages 113-127.

[16] Carpenter, Robert and Jennifer Chu-Carroll. 1998. "Natural Language Call Routing: a Robust, Self-organizing Approach". In *Proceedings of the International Conference on Speech and Language Processing*. Sydney. Volume 5, Pages 2059-2062.

[17] Cattoni, R., M. Federico and A. Lavie, 2001. "Robust Analysis of Spoken Input Combining Statistical and Knowledge-based Information Sources". In Proceedings of Automatic Speech Recognition and Understanding (ASRU-2001), Madonna di Campiglio, Italy, December 2001.

[18] M. Gavaldà, SOUP: A Parser for Real-World Spontaneous Speech. In *Proceedings of the sixth International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy. 2000. Pages 101-110.

[19] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg Memory Based Learner, version 4.3, Reference Guide. ILK Technical Report 02-10. Available from http://ilk.kub.nl/downloads/pub/papers/ilk0210.ps.gz. 2002.

[20] C. Langley, A. Lavie, L. Levin, D. Wallace, D. Gates, and K. Peterson. Spoken Language Parsing Using Phrase-Level Grammars and Trainable Classifiers. In *Proceedings of Workshop on Algorithms for Speech-to-Speech Machine Translation at ACL-02*. Philadelphia, PA. 2002. Pages 15-22.

[21] C. Langley and A. Lavie. Parsing Domain Actions with Phrase-Level Grammars and Memory-Based Learners. in *Proceedings of Eighth International Workshop on Parsing Technologies (IWPT-2003)*. Nancy, France. 2003. Pages 127-136.

[22] L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace and K. Peterson. "Domain Specific Speech Acts for Spoken Language Translation". In *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue (SIGDIAL*-2003), Sapporo, Japan, July 2003.

[23] F. Metze, J. McDonough and H. Soltau. Speech Recognition over NetMeeting Connections. In *Proceedings of EuroSpeech 2001*, Aalborg, Denmark. ISCA. 2001.

[24] A. Lavie, F. Metze, R. Cattoni and E. Costantini. "A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System". In *Proceedings of Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, Philadelphia, PA, July 2002. Pages 121-128.

[25] S. Burger, E. Constantini and F. Pianesi. Communicative Strategies and Patterns of Multimodal Integration in a Speech-to-Speech Translation System. In *Proceedings of MT Summit IX*, New Orleans, LA. 2003. Pages 32-39.

[26] S. Burger, E. Constantini and F. Pianesi. Communicative Strategies in a Speech-to-Speech Translation System. *Intelligenza Artificiale, 1 (1)*, February, 2004. Pages 52-56.

[27] J. Carletta, S. Isard, A. H. Anderson, G. Doherty-Sneddon, A. Isard and J. C. Kowtko. "The Reliability of a Dialogue Structure Coding Scheme". *Computational Linguistics, 23 (1)*, MIT Press. 1997. Pages 13-21.

[28] J. Carletta et al. 1996. HCRC Dialogue Coding Manual", *HCRC Technical Report*, HCRC/TR-82.

[29] A. Lavie et al. Evaluation of the NESPOLE! Showcase-2a System. Deliverable D18. Avialable on NESPOLE! Website at: http://nespole.itc.it