Enhancing the Usability and Performance of Nespole! - a Real-World Speech-to-Speech Translation System

Alon Lavie, Florian Metze, Fabio Pianesi, Susanne Burger, Donna Gates, Lori Levin, Chad Langley, Kay Peterson, Tanja Schultz, Alex Waibel, et al.

Enhancing the Usability and Performance of NESPOLE! — a Real-World Speech-to-Speech Translation System

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

Florian Metze
Interactive Systems Laboratories
Universität Karlsruhe (TH)
Karlsruhe, Germany
metze@ira.uka.de

Fabio Pianesi
ITC-irst
Trento, Italy
pianesi@itc.it

1. INTRODUCTION

NESPOLE! [1] 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; Ancona, Italy – a tele-communications company). The project is funded jointly by the European Commission and the US NSF. Over the past year, we have developed a fully functional showcase of the NESPOLE! system within the domain of travel and tourism[2], and have significantly improved system performance and usability based on a series of studies and evaluations with real users. Our experience has shown that improving translation quality is only one of several important issues that must be addressed in achieving a practical real-world speech-to-speech translation system. This paper describes how we tackled these issues and evaluates their effect on system performance and usability. We focus on three main issues: (1) a study on the usage and utility of multi-modality in the context of multi-lingual communication; (2) assessing system performance under various network traffic conditions and architectural configurations; and (3) an end-to-end evaluation of the demonstration system.

2. THE NESPOLE! SYSTEM

The design principles of the NESPOLE! [4] system were already described in [3]. The 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's NetMeeting® is used to connect between the two parties in real-time.

In the first showcase which we describe in this paper, the scenario is the following: a client user is browsing through the web-pages of APT – the tourism bureau of the province of Trentino in Italy – in search of winter-sport tour-packages in the Trentino region. If more detailed information is desired, the client can click on a dedicated "button" within the web-page in order to establish a video-conferencing connection to a human agent located at APT. 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! server provides two-way speech-to-speech translation between the parties. In the current setup, the agent speaks Italian, while the client can speak English, French or German.

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 very 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. A well-defined 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 software packages.

For example, let us suppose an English-speaking client in the US is connecting to an APT agent in Italy. A connection request from the client's PC (in the US) would be

[1] NESPOLE! – NEgotiation through SPOken Language in E-commerce. See the project web-site at http://nespole.itc.it for further details.
[2] A demonstration of this showcase system will be shown at the HLT-2002 conference

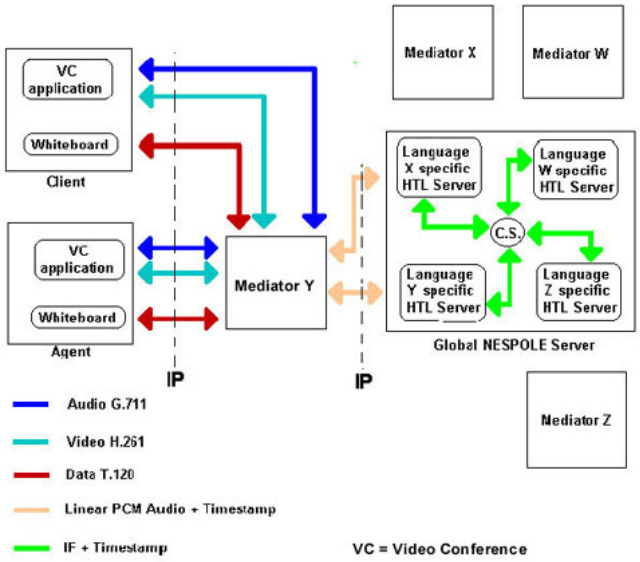

Figure 1: The Nespole! System Architecture

made to the Mediator, which can be physically located anywhere on the net (but practically most likely located either in the US or in Italy). The Mediator establishes a connection over the internet with both an English HLT server and the Italian HLT server (which can again be physically located anywhere on the internet), before calling the agent in Trento. The English HLT server provides English speech recognition, translation from English text into our Interlingua "IF" (Interchange Format) as well as English generation and speech synthesis from IF. The Italian HLT server provides similar functionalities to and from Italian. This basic system design is shown in Figure 1.

The computationally intensive part of speech recognition and translation is done on dedicated server machines, whose nature and location is of no concern to the user. A wide range of client-machines, even portable devices or public information kiosks, are therefore able to run the client software, so that the service can be made available nearly everywhere. The main technical difficulty for VoIP ("Voice over Internet Protocol") applications is coping with adverse internet bandwidth conditions. In order to guarantee real-time communication under insufficient bandwidth conditions, video-conferencing software often drops short segments of speech that were delayed in transport. This, however, can be very detrimental to the performances of speech recognizers [7]. Later in this paper, we present performance statistics, collected with the actual system, that investigate the performance effects of this problem within our system . In order to reduce bandwidth requirements, it is possible to use the NESPOLE! system without video transmission, as we currently do not transmit critical information over this channel.

3. ENHANCING PERFORMANCE AND USABILITY OF THE NESPOLE! SYSTEM

During a preliminary user study conducted in Summer-2001 to evaluate the initial version of our system, several key points emerged that allowed us to improve the system to its current configuration. The resulting "First Showcase" of our system has been demonstrated at several events (e.g. the IST conference in Düsseldorf, Germany, in December 2001) and is also presented at the HLT-2002 demonstration session [6]). The recent improvements allow the users to retain control of the interaction in all situations, and permit a smooth integration of multi-modal and multi-lingual communication. In the following subsections, we describe the main issues we have addressed in the course of developing the current showcase system.

3.1 System Response Time

The standard Video-conferencing setup which we use in the NESPOLE! system allows transmission of both audio and video streams between the participants. In order to limit bandwidth requirements, we currently do not use the video channel and transmit only audio. The system architecture shown in figure 1 contains two different types of Internet connections with different characteristics:

- The "left" IP 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 "right" IP connection between the Mediator and the HLT servers uses TCP over IP in order to achieve lossless communication between the Mediator, which interfaces with the UDP channels, and the translation components. For practical reasons, Mediator and HLT servers in our current system usually run in separate and distant locations, which can introduce some time delays, in addition to the time needed for actual data processing by the HLT components.

Speech-packets transmitted over the H323 protocol can be processed as soon as they arrive without waiting for subsequent packets that contain the remaining speech. The inclusion of "Run-On" processing in the speech recognizers and the optimization of inter-process communication within the HLT servers reduced the answering times of the overall system to approximately three times real-time in recent demonstrations.[3] Our experience indicates that users appreciate the fact that the system provides some feedback on its progress in the form of text messages (i.e. recognition results), even before paraphrasing and translation are complete.

3.2 Adaptation of Automatic Speech Recognition Engines

The Speech Recognition modules of the Nespole! system were developed separately at the different participating sites, using different toolkits, but communicate with the Mediator using a standardized interface. The French and German

[3]For comparison, German Speech recognition runs in about real-time on a standard 1GHz Pentium-III machine using a 12k vocabulary.

	English	French	German	Italian
Vocabulary size	8,000	20,000	12,000	4,000
OOV rate	0.3%		<1%	3.0%
LM training Data	Verbmobil (E), C-Star 550k words	Internet 1,500M words	Verbmobil (D) 500k words	Nespole 1500 sentences
+ adaptation	Nespole	Nespole	Nespole	
Perplexity	33		98	150
Microphone type	head-set	head-set	table-top	head-set
Speaking style	spontaneous	read	spontaneous	read
Ac. training Data	16kHz 90h	G711 recoded 12h	16kHz 65h Verbmobil-II	G711 recoded 11h C-Star
+ adaptation	Upsampling of G711		MLLR 80min. + FSA	
Real-time factor	2.5, 1GHz P-III		1.1, 1GHz P-III	1.8, 650Mhz P-III
Memory consumption	280Mb	200Mb	100Mb	
WER on clean data	19.9%	18% (read speech)	29.8%	31.5%

Table 1: Speech Recognizers Used in the NESPOLE! System

ASR modules are described in more detail in [9, 7]. The German engine was derived from the UKA recognizer developed for the German Verbmobil Task [10].

All systems were derived from existing LVCSR recognizers and adapted to the Nespole! task using less than 2 hours of adaptation data. This data was collected during an initial user-study, in which clients from all countries communicated with an APT agent fluent in their mother tongue through the Nespole! system, but without recognition and translation components in place. Segmentation of input speech is done based on automatic silence detection performed by NetMeeting at the site of the originating audio. The audio is encoded according to the G.711 standard at a sampling frequency of 8kHz. The characteristics of the different recognizers are summarized in Table 1.

The word accuracy rates obtained by the recognition engines for the various languages are presented together with the results of the end-to-end evaluation in Section 4.3.

3.3 Improved Analysis Accuracy

The analysis engines for the various languages handled by our system are developed independently by our research groups and follow different approaches. We describe here some recent novel common aspects of the analysis modules developed for English, German and Italian[4]. The analysis process in our system involves mapping input utterances into an interlingua representation ("IF"), which consists of two main pieces of information: (1) the Domain Action (a speech act and a sequence of concepts); and (2) arguments, consisting of feature-value information. Our analyzers apply two separate stages to extracting these two types of information from an input utterance. Statistical models and classifiers are used for the former, while a knowledge-based approach is employed for the latter. Decomposing the problem of IF construction into two separate and independent sub-tasks has the advantage of allowing specialized techniques to be applied to each sub-task, while reducing the amount of needed human grammar development. The main drawback, however, is that the information extracted by the two separate sub-tasks may be inconsistent when combined together, resulting in illegal IFs. Particular attention has thus been given to the problem of the production of legal IFs, since it has a crucial impact on the practical use of the system. The analyzers for English, German and Italian employ procedures for using the well-formedness constraints defined by the Interlingua formalism in order to select the best scoring Domain Action that is most compatible with the extracted arguments [2].

[4]The French analyzer uses a significantly different approach than the one described here.

3.4 Multi-Modality

The nature of the e-commerce scenario and application in which our system is situated requires that speech-translation be well-integrated with additional modalities of communication and information exchange between the agent and client. Significant effort has been devoted to this issue within the project. The main multi-modal component in the current version of our system is the AeWhiteboard – a special whiteboard, which allows users to share maps and webpages. The functionalities provided by the AeWhiteboard include: image loading, free-hand drawing, area selecting, color choosing, scrolling the image loaded, zooming the image loaded, URL opening, and Nespole! Monitor activation. The most important feature is that each operation the user does is shared with his remote interlocutor, so they can communicate while viewing the same images and drawing on identical-image whiteboards.

Typically, the client asks for information regarding locations, distances between locations, and navigation directions, e.g., how to get from a proposed hotel to the ski slopes. By using the whiteboard, the agent can indicate the locations and draw routes on the map, point at areas, select items, draw connections between different locations using a mouse or an optical pen, and accompany his/her gestures with verbal explanations. Supporting such combined verbal and gesture interactions has required modifications and extensions of both HLT modules and the IF. In summer 2001, we conducted a detailed study to evaluate the effects of multi-modality on the communication effectiveness and usability of our system. The main results of this study are presented in the evaluation section of this paper.

3.5 User interface

Significant attention was also devoted to designing an appropriate front-end user interface for the system, that allows both clients and agents an intuitive and relatively simple control over their communication process.

The user interface display is designed for Windows®and consists of four windows: (1) a Microsoft®Internet Explorer web browser; (2) a Microsoft®Windows NetMeeting videoconferencing application; (3) the AeWhiteboard; and (4) the Nespole Monitor. Using Internet Explorer, the client initiates the audio and video call with an agent of the service provider, by a simple click of a button on the browser page. Microsoft Windows Netmeeting is automatically opened and the audio and video connection is established. The two additional displays – the AeWhiteboard and the Nespole Monitor are also launched at the same time. Client and agent can then proceed in carrying out a dialogue with the help of the speech translation system. For a screen snapshot of these four displays, see [6].

We found it important to visually present aspects of the speech-translation process to the end users. This is accomplished via the Nespole Monitor display. Three textual representations are displayed in clearly identified fields: (1) a transcript of their spoken input (the output from the speech recognizer); (2) a paraphrase of their input – the result of translating the recognized input back into their own language; and (3) the translated textual output of the utterance spoken by the other party. These textual representations provide the users with the capability to identify mistranslations and indicate errors to the other party. A bad paraphrase is often a good indicator of a significant error in the translation process. When a mis-translation is detected, the user can press a dedicated button that informs the other party to ignore the translation being displayed, by highlighting the textual translation in red on the monitor display of the other party. The user can then repeat the turn. The current system also allows the participants to correct speech recognition and translation errors via keyboard input, a feature which is very effective when bandwidth limitations degrade the system performance.

4. EVALUATION

Several different evaluation experiments have been conducted, targeting different aspects of our system: (1) the impact and usability of multi-modality; (2) experiments for assessing the impact of network traffic and the consequences of real packet-loss on system performance; and (3) end-to-end performance evaluations. The database collected during the project and which is being used in the various evaluations is described in [1].

4.1 Experiments on Multi-Modality

During July 2001, we conducted a detailed study to evaluate the effect of multi-modality on the communication effectiveness and usability of our system. The goals of the experiment were to test: (1) whether multi-modality increases the probability of successful interaction, especially when spatial information is the focus of the communicative exchange; (2) whether multi-modality helps reduce ambiguities and disfluencies; and (3) whether multi-modality supports a faster recovery from recognition and translation errors. For these purposes, two experimental conditions were devised: a speech-only condition (SO), involving multilingual communication and the possibility for users to share images; and a multi-modal condition (MM), where users could additionally perform pen-based gestures on shared maps to convey spatial information.

The setting for the experiment was the NESPOLE! scenario described earlier, involving clients searching for winter tourpackage information in the Trentino province. The client's task was to select an appropriate resort location and hotel within the constraints specified a priori concerning the relevant geographical area, the available budget, etc. The agent's task was to provide the necessary information. Novice subjects, previously unfamiliar with the system and task were recruited to play the role of the clients. Subjects wore a head-mounted microphone, using it in a push-to-talk mode, and drew gestures on maps by means of a table-pen device or a mouse. Each subject could only hear the translated speech of the other party (original audio was disabled in this experiment). The experiment was conducted on a total of 28 recorded dialogues, with 14 dialogues each for English and for German clients, and Italian agents in all cases. Each group contained seven SO and seven MM dialogues. The dialogue corpus consisted of 16.5 hours of dialogue length: 8.5 hours of English-Italian, 8 hours of German-Italian. The average duration of dialogues was 35 minutes. On average, a dialogue contained 35 turns, 247 tokens and 97 token types per speaker. The dialogue transcriptions include: orthographical transcription, annotations for spontaneous phenomena and disfluencies, turn information and annotations for gestures. Translated turns were classified into successful, partially successful and non-successful, by comparing the translated turns with the responses they generated. Repeated turns were counted as well.

The analysis of the results indicated that both the SO and MM versions of the system were effective for goal completion: 86% of the users were able to complete the task's goal by choosing a hotel meeting the pre-specified budget and location constraints. This demonstrates that the system is sufficiently adequate for novice users to accomplish the given task with minimal written instructions, a very short initial training on using the whiteboard, and no further assistance during the interaction. The average number of gestures per dialogue was 7.6. Gestures were performed only when spatial information was involved, and this ocurred in only a few of the dialogue segments. The agents performed almost all drawing, with a clear preference for area selections. Almost every gesture followed a dialogue contribution. Overall, few or no deictics were used. We believe that these findings are related to the push-to-talk procedure and to the time needed to transfer gestures across the network. Gestures were always preceded by appropriate verbal cues—e.g., "I'll show you the ice skating rink on the map". This shows that gestures were well integrated in the communication.

Data analysis also indicated that the number of turns, the number of words and the dialogue length were similar across conditions and languages. An analysis of the impact of MM on the language used revealed that there were fewer repeated turns (indicating a better reciprocal understanding of the two parties) and smoother dialogues (with fewer returns to

already discussed topics) under the MM condition. MM also exhibited fewer ambiguous utterances. Furthermore, the ambiguities in MM conditions were often immediately solved by resorting to MM resources. This was not the case in SO, where ambiguous or mis-understood utterances often remained unresolved. The experiment subjects, given the choice between the MM and the SO system, expressed a clear preference for the former. In summary, the results clearly indicate that multi-modal and multi-lingual features were smoothly integrated. Furthermore, MM had a positive effect on the quality of interaction by reducing ambiguity, making it easier to resolve ambiguous utterances, improving the flow of the dialogue, and enhancing the mutual comprehension between the parties.

4.2 Network Traffic Impact

In our various user studies and demonstrations, we have been 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, which rely on the UDP protocol, in order to guarantee real-time human-to-human communication. The communication between the Mediator and HLT servers can, in principle, be within a local network, although we currently run the HLT servers at the sites of the developing partners. This introduces time delays, but no packet loss, due to the use of TCP, in contrast with the UDP used for the H323 connections.

To quantify the influence of UDP packet-loss on system performance, we ran a number of tests between German client installations in the USA (CMU at Pittsburgh) and Germany (UKA at Karslruhe) calling a Mediator in Italy (IRST), which in turns contacted the German HLT server located in Karlsruhe, Germany. The tests were conducted by feeding a high-quality recording of the German development-test set collected at the beginning of the project into a computer set-up for a video-conference, i.e. we replaced the microphone by a DAT recorder (or a computer) playing a tape, while leaving everything else as it would be for sessions with real subjects. In particular, segmentation was based on silence detection performed automatically by NetMeeting. The resulting segments were recognized separately by the HLT servers and the hypotheses concatenated to calculate the WER over the whole dialogue. These tests (a total of more than 16 hours) were conducted at different times of the day on different days of the week, in an attempt to investigate a wide as possible variety of real-life network conditions.

All in all, we were able to run 16 complete tests, resulting in an average word accuracy of 60.4%,[5] with single values in the 63% to 59% range for packet-loss conditions between 0.1% and 5.2%. Higher packet-loss ratios, resulting from generally bad network conditions, usually led to a breakdown of the Client-Mediator or Mediator-HLT server link due to time-out conditions being reached, or the inability to establish a connection at all. We were able, however, to record one dialogue with 21.0% packet loss, which resulted in a word accuracy of 50.3%. This dialogue is very difficult to understand even for humans. From the recorded statistics we conclude that at least for packet-loss ratios below 5%, this number alone is not sufficient to predict word-error rate. For 20% packet-loss, the loss in WER is significant, but we still observe less degradation than reported in [8] on synthetic data. In most cases, our experience indicates that we are likely to face packet-loss ratios below 5%, where there is no clear correlation between packet-loss and word-error rate (WER), and WERs remain at levels close to those experienced under optimal bandwidth conditions.

[5]The word accuracy on the clean 16kHz recording is 71.2%.

Language	WARs	SR Graded (% Acc)
English	61.9%	66.0%
German	63.5%	68.0%
French	71.2%	65.0%
Italian	76.5%	N/A

Table 2: Speech Recognition Word Accuracy Rates and Results of Human Grading (Percent Acceptable) of Recognition Output as a Paraphrase

Language	Transcribed	Speech Rec.
English-to-English	58%	45%
German-to-German	46%	40%
French-to-French	54%	41%
Italian-to-Italian	61%	48%

Table 3: Monolingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input

4.3 End-to-End System Evaluation

In December 2001, we conducted a large scale multi-lingual end-to-end translation evaluation of the Nespole first-showcase system. For each of the three language pairs (English-Italian, German-Italian and French-Italian), four test dialogues that were not previously seen by the system developers were used to evaluate the performance of the translation system. The dialogues included two scenarios: one covering winter ski vacations, the other about summer resorts. One or two of the dialogues for each language contained multi-modal expressions. The test data included a mixture of dialogues that were collected mono-lingually prior to system development (both client and agent spoke the same language), and data collected bilingually (during the July 2001 MM experiment), using the actual translation system. This mixture of data conditions was intended primarily for comprehensiveness and not for comparison of the different conditions.

We performed an extensive suite of evaluations on the above data. The evaluations were all end-to-end, from input to output, not intended to assess individual modules or components. We performed both mono-lingual evaluation (where the generated output language was the same as the input language), as well as cross-lingual evaluation. For cross-lingual evaluations, client utterances were evaluated on translation from English German and French to Italian, and agent utterances were evaluated from Italian to each of the three languages. We evaluated on both textual manually transcribed input as well as on input from actual speech-recognition of the original audio. We also graded the speech recognized output as a "paraphrase" of the transcriptions, to measure

Language	Transcribed	Speech Rec.
English-to-Italian	55%	43%
German-to-Italian	32%	27%
French-to-Italian	44%	34%
Italian-to-English	47%	37%
Italian-to-German	47%	31%
Italian-to-French	40%	27%

Table 4: Cross-lingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input

the levels of semantic loss of information due to recognition errors. Speech recognition word accuracies and the results of speech graded as a paraphrase appear in Table 2. Translations were graded by multiple human graders at the level of Semantic Dialogue Units (SDUs). For each data set, one grader first manually segmented each utterance into SDUs. All graders then used this segmentation in order to assign scores for each SDU present in the utterance. We followed the three-point grading scheme previously developed for the C-STAR consortium, as described in [5]. Each SDU is graded as either "Perfect" (meaning translated correctly and output is fluent), "OK" (meaning is translated reasonably correct but output may be disfluent), or "Bad" (meaning not properly translated). We calculate the percent of SDUs that are graded with each of the above categories. "Perfect" and "OK" percentages are also summed together into a category of "Acceptable" translations. Average percentages are calculated for each dialogue, each grader, and separately for client and agent utterances. We then calculated combined averages for all graders and for all dialogues for each language pair.

Table 3 shows the results of the monolingual end-to-end translation for the four languages, and Table 4 shows the results of the cross-lingual evaluations. The results indicate acceptable translations in the range of 27–43% of SDUs (interlingua units) with speech recognized inputs. While this level of translation accuracy cannot be considered impressive, our user studies and system demonstrations indicate that it is already sufficient for achieving effective communication with real users. We expect performance levels to reach a range of 60–70% within the next year of the project.

5. ACKNOWLEDGMENTS

The research work reported here was supported by the National Science Foundation under Grant number 9982227 and the European Union under Grant number IST 1999-11562 as part of the joint EU/NSF MLIAM research initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the EU or the NSF.

6. ADDITIONAL AUTHORS

S. Burger, D. Gates, L. Levin, C. Langley, K. Peterson, T. Schultz, A. Waibel and D. Wallace (Carnegie Mellon; Pittsburgh, PA, USA; email: {sburger|dmg|lsl|clangley|kay+|tanja+|ahw|dorcas}@cs.cmu.edu), and J. McDonough, and H. Soltau (UKA, Karlsruhe, Germany; email: {jmcd|soltau}@ira.uka.de), and R. Cattoni, G. Lazzari, N. Mana, and E. Pianta (ITC-irst, Trento, Italy; email: {cattoni|lazzari|mana|pianta}@itc.it), and E. Costantini (University of Trieste, Italy; email: costanti@psico.univ.trieste.it), and L. Besacier, H. Blanchon, and D. Vaufreydaz (CLIPS; Grenoble, France; e-mail: {Laurent.Besacier|Herve.Blanchon|Dominique.Vaufreydaz}@imag.fr), and L. Taddei (AETHRA; Ancona, Italy; e-mail: l.taddei@aethra.it).

7. REFERENCES

[1] S. Burger, L. Besacier, P. Coletti, F. Metze, and C. Morel. The NESPOLE! VoIP Dialogue Database. In *Proc. EuroSpeech 2001*, Aalborg, Denmark, 2001. ISCA.

[2] R. Cattoni, M. Federico, and A. Lavie. Robust Analysis of Spoken Input combining Statistical and Knowledge-based Information Sources. In *Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, Madonna di Campiglio, Italy, December 2001.

[3] A. Lavie, F. Pianesi, and 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.

[4] G. Lazzari. Spoken translation: challenges and opportunities. In *Proc. ICSLP 2001*, Beijing, China, October 2001.

[5] L. Levin, D. Gates, A. Lavie, F. Pianesi, D. Wallace, T. Watanabe, and M. Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings of NAACL-2000 Workshop On Interlinguas and Interlingual Approaches*, Seattle, WA, 2000. ACL.

[6] 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)*, March 2002.

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

[8] B. Milner and S. Semnani. Robust Speech Recognition over IP Networks. In *Proc. ICASSP 2001*, Salt Lake City, USA, May 2001.

[9] S. Rossato, H. Blanchon, and L. Besacier. Evaluation of a Speech to Speech Translation System: French Experience in the NESPOLE! Project. In *Submitted to Proc. 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, August 2002.

[10] H. Soltau, T. Schaaf, F. Metze, and A. Waibel. The ISL Evaluation System for Verbmobil - II. In *Proc. ICASSP 2001*, Salt Lake City, USA, May 2001.