The NESPOLE! Speech-to-Speech Translation System

F. Metze, J. McDonough, H. Soltau, and A. Waibel*
Interactive Systems Laboratories
Universität Karlsruhe (TH)
Karlsruhe, Germany
{metze|jmcd|soltau|waibel}@ira.uka.de

A. Lavie, S. Burger, C. Langley, L. Levin, and T. Schultz
Carnegie Mellon University
Pittsburgh, PA, USA
{alavie|sburger|clangley|lsl|tanja+}@cs.cmu.edu

F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, and E. Pianta
ITC-irst
Trento, Italy
{pianesi|cattoni|lazzari|mana|pianta}@itc.it

1. INTRODUCTION

This paper describes the first showcase of the NESPOLE![1] system. NESPOLE! is a speech-to-speech machine translation system 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 laboratories (IRST in Trento, Italy; ISL at Universität Karlsruhe (TH) in Germany; 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.

The main goal of NESPOLE! is to advance the state-of-the-art of speech-to-speech translation in realistic scenarios and involving naive users. The first showcase presented in this demonstration involves an English–, French–, or German–speaking client enquiring about winter-sports possibilities in the Trentino region of the Italian Alps via a NetMeeting® connection. His or her questions are answered by an Italian-speaking agent at APT, while the NESPOLE! system provides speech-to-speech translation and a multi-modal Whiteboard, which users can use to point to shared web-sites or draw on shared maps, therefore enhancing the oral communication with extra capabilities.

This paper is organized as follows: we will first present a general system description which covers the current hardware setup as well as the design principles of the NESPOLE! system in general. We will then present the user interface, and the results from two system evaluations, one with respect to system performance under different network conditions and an end-to-end evaluation. The references at the end of this paper serve as pointers to further information about the NESPOLE! system.

[1]NESPOLE! – NEgotiation through SPOken Language in E-commerce.
*Also at Carnegie Mellon University, Pittsburgh, PA.

Figure 1: A user during a NESPOLE! video-conferencing session.

2. SYSTEM DESCRIPTION

The NESPOLE! system uses a client-server architecture to allow a common user, who is browsing web-pages on the Internet, to connect seamlessly to a speech-to-speech translation service using for example Microsoft's NetMeeting® software. In the current showcase, a user browses the web-pages of the Trentino region in Italy for winter-sports possibilities and, as he doesn't find all the information he wants, clicks on a button provided on the web-page to establish a video-conferencing[2] connection to a human "agent" of the Trentino tourist board (see figure 1). He can then ask natural-language questions to the agent, which the NESPOLE! server will translate and vice-versa. Currently, the

[2]The use of the video-conferencing feature is optional and can be dropped to reduce bandwidth, as the system functionality is fulfilled entirely by the data and audio streams.

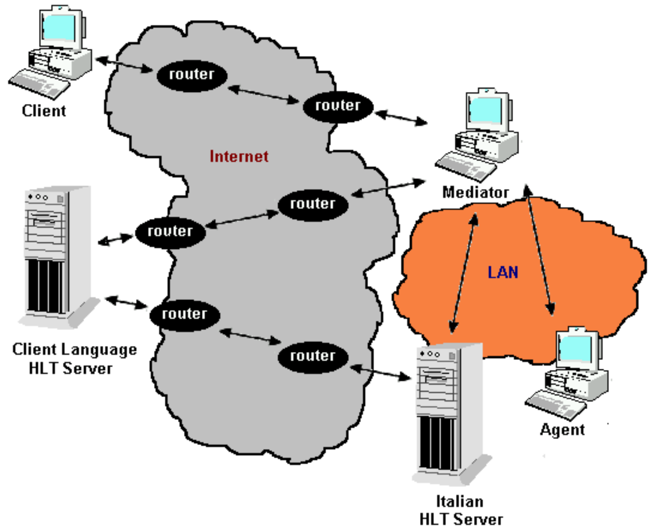

Figure 2: The distributed NESPOLE! system architecture.

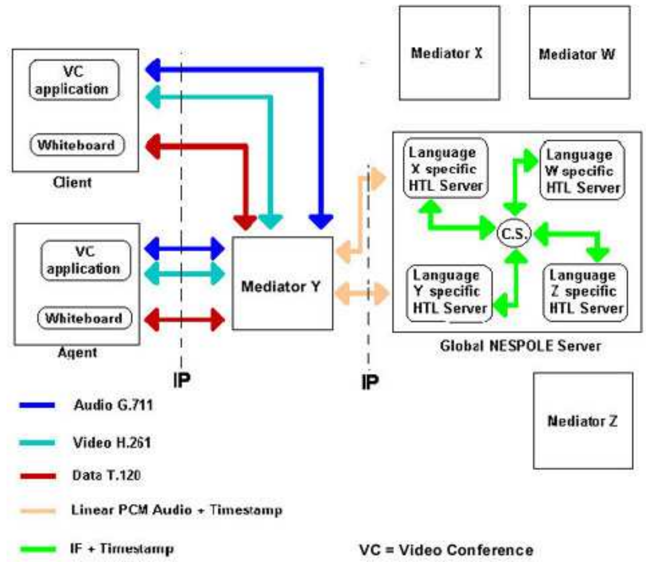

Figure 3: The logical view of the NESPOLE! system.

agent speaks Italian, while the client can speak either English, French or German.

The Interlingua-based translation system covers the activities of planning and scheduling winter holidays and similar activities in the Trentino region. By using NESPOLE!, customers can be served in several languages without the need to employ agents capable of speaking all of them. Additionally, the NESPOLE! system incorporates a special Whiteboard with multi-modal capabilities, allowing the agent and the client to share maps or web-pages.

2.1 Hardware requirements

The system requires no special hardware on the client's side, except a standard PC or portable device with microphone and loudspeakers or headsets as well as an Internet connection with a bandwidth of about 64kbit/s. We have for example demonstrated the system on a laptop running Windows 2000, connected to the Internet via a wireless LAN link.

The hardware setup used within the NESPOLE! system is shown in figure 2. The client connects to a special server, the so-called "Mediator" machine, which then in turn establishes connections to the so-called "HLT-servers", which provide the ASR and MT capabilities. The Mediator also runs under Windows, while the HLT servers run on different flavours of Linux on Intel PCs or Unix workstations. The IP connections between the Mediator computer and the agent and client use the H323 video-conferencing standard, which is based on UDP for the audio stream. Data is transmitted via TCP. This means that there will be little time delay during transmission, which is important for human-to-human communication, but short segments of speech can be lost during transmission. The links between the Mediator and the HLT servers use TCP. The effect of packet-loss on the demo will be discussed in the system evaluation section. The logical system design is shown in figure 3.

The system complexity is hidden from the user, as he only communicates with the Mediator computer. Currently, we usually run the Mediator at IRST in Trento, while the agent is being called at APT, also in Trento. The HLT servers, which provide speech recognition, translation and synthesis, run at the locations of the participating partners, i.e. at the Universities in Trento, Pittsburgh, Grenoble and Karlsruhe.

The chosen design allows for maximum flexibility during usage: Mediators and HLT-servers can be run in several locations, so that the optimal configuration given the current user location and network traffic can be chosen at runtime. The computationally intensive part of speech recognition and translation is done on dedicated server machines. The client machine can therefore be very "thin", so that the service is available nearly everywhere and to everyone, including mobile devices and public information kiosks.

2.2 Software

A complete call through the system starts by a client requesting a NetMeeting connection with the Mediator. The Mediator identifies the client's native language by an entry in the "Name" field, which is transmitted during calls and will then contact the HLT server for the client's language, the HLT server for the agent's language (currently Italian), to check the availability of translation services and try to establish a NetMeeting connection to an agent. It will only accept the call from the client, once these three required connections have been established. Speech from the client will be received by the Mediator, forwarded to the respective HLT server, which will in turn perform speech recognition and analysis into a language-independent "Interchange Format" (IF) [6], which is then transmitted to the HLT server associated with the agent, where text will be generated from the IF. This text string will then be synthesized and the resulting audio is transmitted to the agent via the Mediator. Multi-modal gestures, such as drawing on a map or video data is transmitted directly between the communication partners.

The speech recognition and translation components have been developed by the participating partners during the project. Further information on the implementation of these modules can be found in the references at the end of this paper, summarized in section 5. In NESPOLE! we use an Interlingua-based translation approach, because this allows us to easily expand the system to other languages thanks to the star-shaped architecture evident in figure 3, and incorporate an user feed-back loop by generating a paraphrase in the user's own language (useful for error correction).

System response time is highly variable due to the uncertain and varying network conditions. The speech recognition components use run-on recognition, i.e. recognition starts as soon as the first packets of data arrive, and run approximately real-time (German) or less than 3 times real-time (English) on standard 1GHz Pentium-III PCs running Linux. Depending on network conditions, text representations of speech recognition or translation, as discussed in the following section on the user interface, can be available in less than one second after a subject stopped speaking. Under bad network conditions, the same process can however take several seconds, too.

3. USER INTERFACE

Significant attention was therefore 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 Windows®-based and consists of four windows:

- the Microsoft® Internet Explorer
- the Microsoft® Windows NetMeeting
- the AeWhiteboard
- the Nespole Monitor

These windows can be seen in figure 4. Using Internet Explorer, the user activates the audio and video call with an agent who can help him and give him the answers and the details he needs: all the user has to do is to click a button on the browser page and automatically Microsoft Windows NetMeeting is opened and the audio and video connection goes up.

We found it important to visually present aspects of the speech-translation process to the end user. This is accomplished via the Nespole Monitor display. Three textual representations are displayed in clearly identified fields, shown in figure 5:

1. a transcript of the spoken utterance (the output from the speech recognizer);
2. a paraphrase of the utterance – the result of translating the recognized input back into the client's own language;
3. the textual translation of the utterance spoken by the other party.

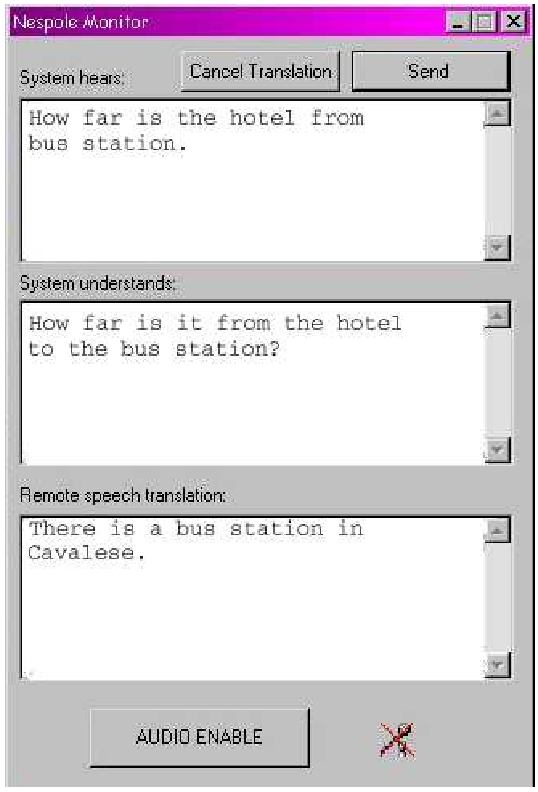

Figure 5: The NESPOLE! monitor window.

These textual representations provide the users with the capability to identify wrong translations and indicate errors to the other party. A bad paraphrase is often a good indicator of a significant error in the translation process. When an inaccurate 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. The "Monitor" window, which allows such interaction, is shown in detail in figure 5.

The AeWhiteboard is provided to improve the quality and the clarity of the conversation through multi-modal tools: it gives to the user the possibility to share with his remote interlocutor an image or a user's free-hand drawing; this is realized in a very simple and intuitive way through a user interface that follows the standards of Windows applications: there is a menu, a tool-bar, and a status bar where the user can read system, state, and button functionality explanations.

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 com-

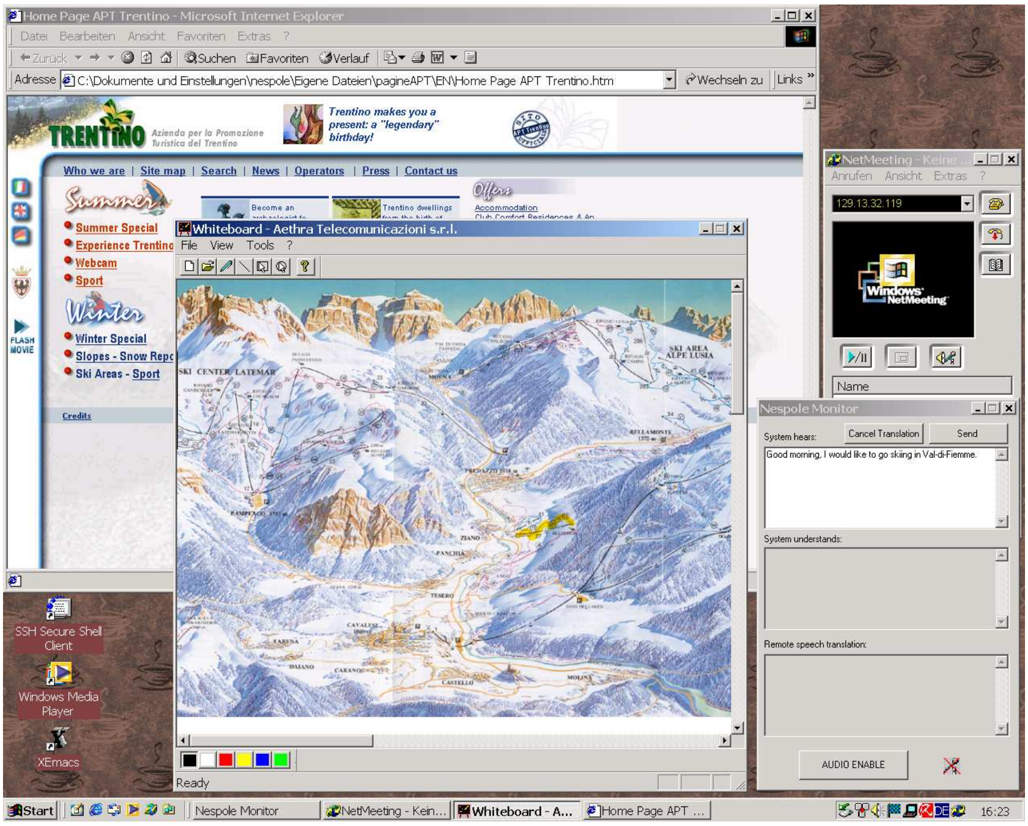

Figure 4: The user's view of the NESPOLE! system.

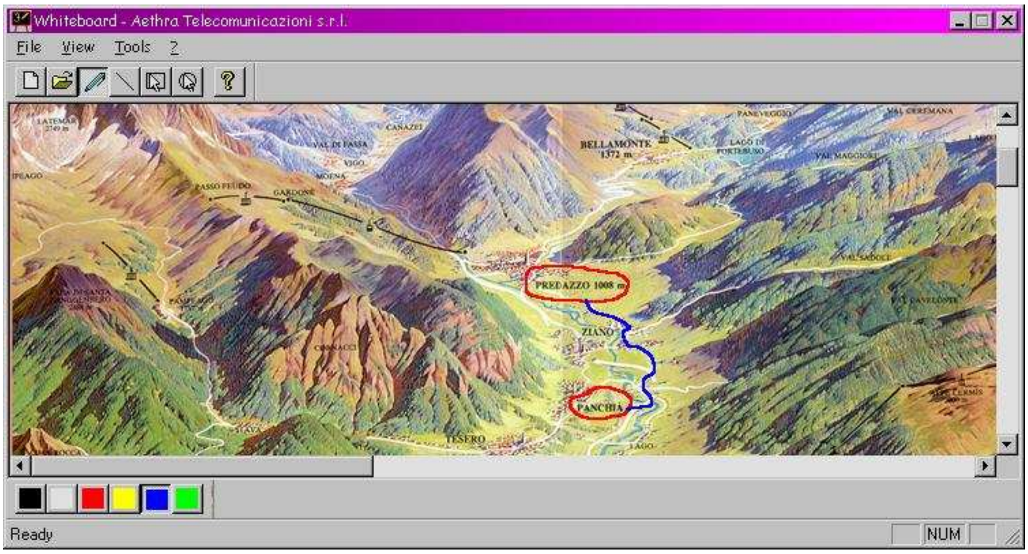

Figure 6: The AeWhiteboard in use.

municate while viewing the same images and drawing on identical-image whiteboards. An example of the AeWhiteboard being used to highlight two towns and hiking path between them is shown in figure 6. This figure is taken from an extension of the system, which we currently develop, that can also handle summer activities.

Typically, the client would ask for distances and directions from a proposed location, say a hotel, to another, e.g., ski slopes. By using the White-Board, the agent can indicate the locations and draw routes on the map, accompanying gestures with verbal explanations. The front-end of the NESPOLE! system allows the participants to communicate entirely with speech and use gestures whenever they want.

4. SYSTEM EVALUATION

Different evaluation sessions have been conducted, targeting different aspects:

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 the system's 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 [2]. The evaluation concerning Multi-Modality is presented in the accompanying HLT2002 poster presentation [4]. In this work we will present the results from Network Traffic impact and the End-to-End evaluation.

4.1 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 VoIP 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. For demonstration purposes, we can of course circumvent the problem by positioning the three components involved (client, agent, mediator) within close proximity, this is however not a realistic scenario.

To quantify the influence of UDP packet-loss on real-world system performance, we ran a number of tests between German client installations in the USA (CMU at Pittsburgh) and Germany (UKA at Karlsruhe) calling a Mediator in Italy (IRST), which in turns contacted the German HLT server located in Karlsruhe. 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 automatically performed by NetMeeting. The thus produced 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.

All in all, we were able to run 16 complete tests, resulting in an average word accuracy of 60.4%,[3] 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. These results are presented in graphical from in figure 7. We were however able 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, which we present in a paper describing the Nespole! demo also in these proceedings, 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 practical use, one is, according to our experience, likely to deal with packet-loss ratios below 5%, where there is no clear correlation between packet-loss and word-error rate.

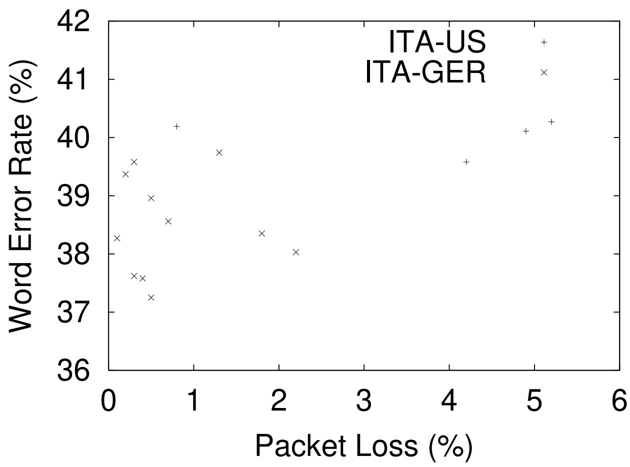

Figure 7: Influence of packet-loss on word accuracy of the German Nespole! recognizer.

Generally, we conclude from these experiments, that packet-loss influences the performance of the system, but there is no clear and drastic influence of packet-loss on ASR performance for the conditions we observed on our experiments. However, because lost packets can also occur during the transmission of synthesized voice, and the original voice is also transmitted between Agent and Client, the user is easily disturbed by the chirping associated with lost packets. A full End-to-End evaluation under all types of network conditions however is beyond the scope of the presented work.

4.2 End-to-End 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

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

(English-Italian, German-Italian and French-Italian), four unseen 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 dialogues 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. The evaluation was conducted by human graders, who manually segmented the dialogues into "Semantic Dialogue Units" (SDUs) and then assigned scores to every SDU present in the utterance. This scoring scheme is discussed in [6].

Language	Transcribed	ASR hypotheses
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 1: Cross-lingual End-to-End Translation Results ("Fraction of Acceptable SDUs (Interlingua Units)") on Transcribed Input and Hypotheses from Speech Recognition.

The results of the cross-lingual evaluation are summarized in table 1. Our results indicate that between 27% and 43% of Interlingua Units (i.e. translation concepts such as "Request to reserve 1 room for 2 persons and 1 week" or "Enquire distance between hotel and bus-stop") have been translated correctly. 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 between real users, especially with the multi-modal capabilities of the present system.

5. FURTHER INFORMATION

The NESPOLE! project [5] has already lead to a number of publications on speech recognition [7, 10, 1] and Interlingua-based speech-to-speech translation [6, 3, 9]. The NESPOLE! database is described in [2]; as the system is currently regularly demonstrated we are still collecting data under a number of different conditions. An accompanying poster presentation is also appearing in these proceedings [4]. The project web-site can be found at `http://nespole.itc.it`.

6. 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.

7. ADDITIONAL AUTHORS

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`).

8. REFERENCES

[1] L. Besacier, H. Blanchon, Y. Fouquet, J. Guilbaud, S. Helme, S. Mazenot, D. Moraru, and D. Vaufreydaz. Speech Translation for French in the NESPOLE! European Project. In *Proc. EuroSpeech 2001*, Aalborg, Denmark, 2001. ISCA.

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

[3] 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, 12 2001.

[4] A. Lavie, L. Besacier, F. Metze, F. Pianesi, and al. Enhancing the Usability and Performance of NESPOLE! - a Real-World Speech-to-Speech Translation System. In *Proc. HLT 2002*, San Diego, CA, 3 2002.

[5] 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.

[6] L. Levin, D. Gates, F. Pianesi, D. Wallace, T. Watanabe, and M. Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proc. AMTA-SIG-IL Workshop On Interlinguas and Interlingual Approaches*, Seattle, WA, 2000. AMTA.

[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, 5 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, 8 2002.

[10] D. Vaufreydaz, L. Besacier, C. Bergamini, and R. Lamy. From generic to task-oriented speech recognition: French experience in the NESPOLE! European project. In *Proc. ITRW Workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001. ITRW.