Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 121-128.
Association for Computational Linguistics.

A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System

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

Florian Metze
University of Karlsruhe, Germany

Roldano Cattoni
ITC-irst, Trento, Italy

Erica Costantini
University of Trieste, Trieste, Italy

Abstract

Performance and usability of real-world speech-to-speech translation systems, like the one developed within the NESPOLE! project, are affected by several aspects that go beyond the pure translation quality provided by the underlying components of the system. In this paper we describe these aspects as perspectives along which we have evaluated the NESPOLE! system. Four main issues are investigated: (1) assessing system performance under various network traffic conditions; (2) a study on the usage and utility of multi-modality in the context of multilingual communication; (3) a comparison of the features of the individual speech recognition engines, and (4) an end-to-end evaluation of the system.

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 two years, we have developed a fully functional showcase of the NESPOLE! system within the domain of travel and tourism, 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 four main issues: (1) assessing system performance under various network traffic conditions and architectural configurations; (2) a study on the usage and utility of multi-modality in the context of multi-lingual communication; (3) a comparison of the features of the individual speech recognition engines, and (4) an end-to-end evaluation of the demonstration system.

[1] NESPOLE! – NEgotiation through SPOken Language in E-commerce. See the project web-site at http://nespole.itc.it for further details.

2 The NESPOLE! System

The NESPOLE! system (Lazzari, 2001) 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 is browsing through the web-pages of APT – the tourism bureau of the province of Trentino in Italy – in search of 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.

2.1 System Architecture

The NESPOLE! system architecture is shown in Figure 1. 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. Further details of the design principles of the system are described in (Lavie et al., 2001).

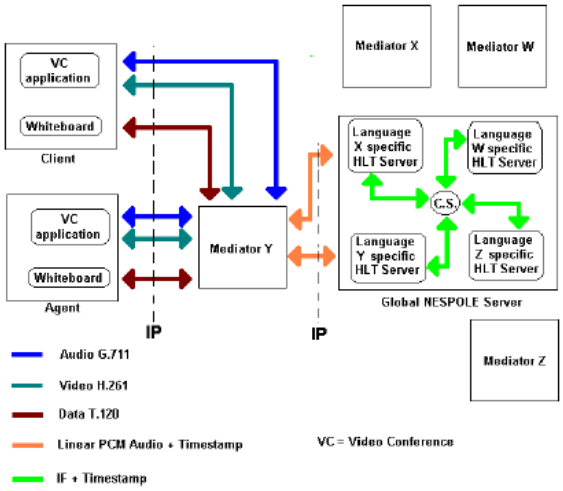

Figure 1: *The Nespole! System Architecture*

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 system architecture shown in Figure 1 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 current 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.

2.2 User interface

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 (Metze et al., 2002).

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 mis-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 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.

3 Multi-Perspective Evaluations

Several different evaluation experiments have been conducted, targeting different aspects of our system: (1) the impact of network traffic and the consequences of real packet-loss on system performance; (2) the impact and usability of multi-modality; (3) a comparison of the features of the various speech recognition engines, developed independently for different languages with different techniques; and (4) end-to-end performance evaluations. The data used in the evaluations is part of a database collected during the project (Burger et al., 2001).

3.1 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 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, but potentially lossy, human-to-human communication. This can potentially be very detrimental to the performance of speech recognizers (Metze et al., 2001). 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, rather than the UDP used for the H323 connections.

To quantify the influence of UDP packet-loss on system performance, we ran a number of tests with German client installations in the USA (CMU at Pittsburgh) and Germany (UKA at Karlsruhe) calling a Mediator in Italy (IRST), which in turn contacted the German HLT server located at UKA. The tests were conducted by feeding a high-quality recording of the German

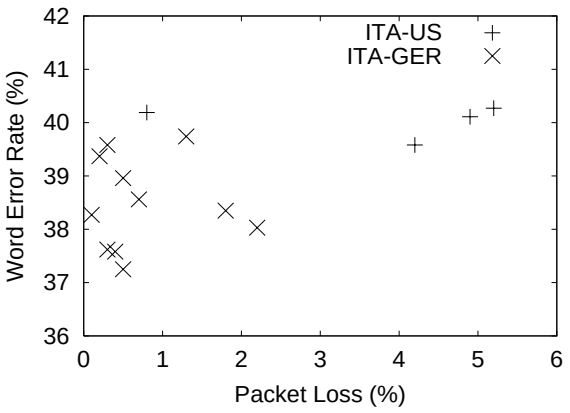

Figure 2: *Influence of packet loss on word accuracy of the German Nespole! recognizer*

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. Each test consisted of several dialogues, lasting about an hour. 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.

We were able to run 16 complete tests, resulting in an average word accuracy of 60.4%,[2] with single values in the 63% to 59% range for packet-loss conditions between 0.1% and 5.2%. The results of these tests are presented in graphical from in Figure 2. On a couple of occasions we experienced abnormally bad network conditions for short periods of time. These 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 full test with 21.0% packet loss, which resulted in a word accuracy of 50.3%. These dialogues are very difficult to understand even for humans.

Our conclusion from the packet loss experiment is that our speech recognition engine is relatively robust to packet loss rates of up to 5%, since there is no clear degradation in the word accuracy of the recognizer as a function of packet loss rate (in this range). This is very good news, since our experience indicates that packet loss rates of over 5% are quite rare under normal network traffic conditions. For 20% packet-loss, the increase in WER is significant, but the degradation is less severe than that reported in (Milner and Semnani, 2000) on synthetic data. We suspect that this is due to the non-random distribution of lost packets.

The tests described above were the first phase of our research on the impact of network traffic on system performance. We are currently in the process of conducting several further experimental investigations concerning different conditions in which the system may run:

Transmission of video in addition to audio through the video-conferencing communication channel: in this case we expect a substantial increase in UDP packet-loss rates due to audio and video competing for the network bandwidth over the H323 connections. It is not clear, however, how this competition takes place in practice and what are the resulting repercussions on the audio quality (and consequently on the recognizers' performance).

The use of low-bandwidth network connections (such as standard 56Kbps modems): This is the most common network scenario for real client users using a home installed computer. We are currently exploring how the bandwidth limitations in this setting affect audio quality and system usability. In low bandwidth conditions, NetMeeting supports encoding the speech with the G.723 codec, which can consume a much lower bandwidth (less then 6.4Kbps) compared to the G.711 codec (64Kbps), which we currently use in our system. We are in the process of testing the G.723 codec within our system. Preliminary results indicate that the recognizers used in the NESPOLE! system are quite robust with respect to this new front-end processing.

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

3.2 Experiments on 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 web-pages. 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 of the whiteboard is that each gesture performed by a user is mirrored on the whiteboard of the other user. Both users communicate while viewing the same images and annotated whiteboards.

Typically, the client asks for spatial information regarding locations, distances, and navigation directions (e.g., how to get from a 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.

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 mis-communications 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 sharing of images; and a multi-modal condition (MM), where users could additionally convey spatial information by pen-based gestures on shared maps.

The setting for the experiment was the scenario described earlier, involving clients searching for winter tour-package information in the Trentino province. The client's task was to select an appropriate resort location and hotel within the specified constraints 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). 28 dialogues were collected, with 14 dialogues each for English and for German clients, and Italian agents in all cases. Each group contained 7 SO and 7 MM dialogues. 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 unsuccessful by comparing the translated turns with the responses they generated. Repeated turns were also counted.

The average duration of dialogues was 35 minutes (35.8 for SO and 35.5 for MM). On average, a dialogue contained 35 turns, 247 tokens and 97 token types per speaker. Average values and variance of all measures are very similar for agents and clients and across conditions and Languages. ANOVA tests ($p=0.05$) on the number of turns and the number of spontaneous phenomena and disfluencies, agents and customers separately, did not produce any evidence that modality or language affected these variables. Hence the spoken input is homogeneous across groups. Details on the experimental database collected and the various statistical analyses performed appear in (Costantini et al., 2002). 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.

In the MM dialogues, there were 7.6 gestures per dialogue on average. The agents performed almost all gestures (98%), with a clear preference for area selections (61% of total gestures). Most gestures (79%) followed a dialogue contribution; none of the gestures were performed during speech. 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: agents often preceded gestures with appropriate verbal cues e.g., "I'll show you the hotel on the map", in order to notify the other party of an upcoming gesture. These verbal cues indicate that gestures were well integrated in the communication.

We found significant differences between the SO and MM dialogues in terms of unsuccessful and repeated turns, particularly so in the spatial segments of the dialogues. In the English-Italian dialogues the MM dialogues contained 19% unsuccessful turns versus 30% for the SO dialogues. For German-Italian dialogues we found 18% in MM versus 31% in SO. English-Italian MM dialogues contained 11% repeated turns versus 17% for SO. For German-Italian dialogues repeated turns amounted to 18% for MM versus 23% for SO. In addition we found smoother dialogues under MM condition, with fewer returns to already discussed topics for MM (one return every 19 turns in SO versus one return every 31 turns in MM). MM also exhibited a lower number of dialogue segments containing identifiable misunderstandings between the two parties (one such segment in each of 3 of the MM dialogues, versus a total of seven such segments in the SO dialogues – one dialogue with 3 segments, one with two, and a third with a single segment of miscommunication). Furthermore, the misunderstandings in MM conditions were often immediately solved by resorting to MM resources, while in case of SO ambiguous or mis-understood subdialogues often remained unresolved. Finally, the experiment subjects, given the choice between the MM and the SO system, expressed a clear preference for the former. In summary, we found strong supporting evidence that multimodality has a positive effect on the quality of interaction by reducing ambiguity, making it easier to resolve ambiguous utterances and to recover from system errors, improving the flow of the dialogue, and enhancing the mutual comprehension between the parties, in particular when spatial information is involved.

3.3 Features 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 ASR modules are described in more detail in (Vaufreydaz et al., 2001; Metze et al., 2001). The German engine was derived from the UKA recognizer developed for the German Verbmobil Task (Soltau et al., 2001).

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 of the recognizers are presented in Section 3.4.

3.4 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 previ-

	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	C-Star 100k words
+ adaptation	Nespole	none	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	Up-sampling 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	100Mb
WER on clean data	19.9%	28%	29.8%	31.5%

Table 1: *Features of the Speech Recognition Engines*

Language	WARs	SR Graded (% Acc)
English	61.9%	66.0%
German	63.5%	68.0%
French	71.2%	65.0%
Italian	76.5%	70.6%

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*

ously unseen test dialogues 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 assessing individual modules or components. We performed both mono-lingual evaluation (where generated output language was the same as the input language), as well as cross-lingual evaluation. For cross-lingual evaluations, translation from English German and French to Italian was evaluated on client utterances, and translation from Italian to each of the three languages was evaluated on agent utterances. We evaluated on both manually transcribed input as well as on actual speech-recognition of the original audio. We also graded the speech recognized output as a "paraphrase" of the transcriptions, to measure 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 seg-

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*

mented 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 (Levin et al., 2000). 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.

Acknowledgements

Additional Authors: S. Burger, D. Gates, C. Langley, K. Laskowski, L. Levin, K. Peterson, T. Schultz, A. Waibel, D. Wallace, Carnegie Mellon University; J. McDonough, H. Soltau, University of Karlsruhe, Germany; G. Lazzari, N. Manna, F. Pianesi, E. Pianta, ITC-irst, Trento, Italy; L. Besacier, H. Blanchon, D. Vaufreydaz, Université Joseph Fourier, Grenoble, France; L. Taddei, AETHRA, Ancona, Italy.

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

References

Susanne Burger, Laurent Besacier, Paolo Coletti, Florian Metze, and Celine Morel. 2001. The NESPOLE! VoIP Dialogue Database. In *Proc. EuroSpeech 2001*, Aalborg, Denmark. ISCA.

Erica Costantini, Susanne Burger, and Fabio Pianesi. 2002. Nespole!'s multilingual and multimodal corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Grand Canary Island, Spain, June. To appear.

Alon Lavie, Fabio Pianesi, and al. 2001. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications. In *Proc. of the HLT2001*, San Diego, CA. ACM.

Gianni Lazzari. 2001. Spoken translation: challenges and opportunities. In *Proc. ICSLP 2001*, Beijing, China, 10.

Lori Levin, Donna Gates, Fabio Pianesi, Donna Wallace, Takeshi Watanabe, and Monika Woszczyna. 2000. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings NAACL-2000 Workshop On Interlinguas and Interlingual Approaches*, Seattle, WA. AMTA.

Florian Metze, John McDonough, and Hagen Soltau. 2001. Speech Recognition over NetMeeting Connections. In *Proc. EuroSpeech 2001*, Aalborg, Denmark. ISCA.

Florian Metze, John McDonough, Hagen Soltau, Alex Waibel, Alon Lavie, Susan Burger, Chad Langley, Kornel Laskowski, Lori Levin, Tanja Schultz, Fabio Pianesi, Roldano Cattoni, Gianni Lazzari, Nadia Mana, Emanuele Pianta, Laurent Besacier, Herve Blanchon, Dominique Vaufreydaz, and Loredana Taddei. 2002. The NESPOLE! Speech-to-Speech Translation System. In *Proc. HLT 2002*, San Diego, CA, 3.

Ben Milner and Sharam Semnani. 2000. Robust Speech Recognition over IP Networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP-00)*, Istanbul, Turkey, June.

Hagen Soltau, Thomas Schaaf, Florian Metze, and Alex Waibel. 2001. The ISL Evaluation System for Verbmobil - II. In *Proc. ICASSP 2001*, Salt Lake City, USA, 5.

D. Vaufreydaz, L. Besacier, C. Bergamini, and R. Lamy. 2001. presented at ISCA ITRW Workshop on Adaptation Methods for Speech Recognition, August. Sophia-Antipolis, France.