The NESPOLE ! Multimodal Speech-to-Speech Translation System: User Based System Improvements

Susannne Burger
Interactive Systems Laboratories
CMU, 5000 Forbes Avenue,
Pittsburgh , PA 15213, USA
+1 412 268 4399
sburger@cs.cmu.edu

Erica Costantini
ITC-Irst
Via Sommarie, Povo
38050, Trento, Italy
+39 0461 314587
costante@itc.it

Fabio Pianesi
ITC-Irst
Via Sommarie, Povo
38050, Trento, Italy
+39 0461 314570
pianesi@itc.it

ABSTRACT
This work discusses the results of two user studies aiming to evaluate the NESPOLE! speech-to-speech translation system, which provides for multilingual and multimodal communication in the tourism and in the medical domain, allowing users to interact through the Internet by sharing maps, web-pages and pen-based gestures. The purpose is to investigate the overall effectiveness of the combination of multimodality and multilinguality in respect to a speech-only system. Specifically, we will examine the improvement in usability offered by the multimodal system, which is gauged by evaluating the interaction between novice users and the system at several stages of development. Additionally, we will discuss how data collection contributed to system improvement.

Keywords
User study, multimodality, machine translation.

INTRODUCTION
NESPOLE! (Negotiation through SPOken Language in E-commerce) is a jointly EU/NSF funded project designed to provide fully functional speech-to-speech capabilities within real-world settings of common users involved in e-commerce applications. The design principles of the NESPOLE! system are described in [1]. The NESPOLE! system uses a client-server architecture to allow a user, who is initially browsing through the web pages of a service provider on the Internet, to connect seamlessly to a human agent who speaks another language, and provides speech-to-speech translation (STST) service. Commercially available PC video-conferencing technology is used to connect between the two parties in real-time. The interface description is available in [3]. The languages addressed are Italian, German, English and French. The scenario for the first two showcases (1 and 2a) is the tourism domain and involves an Italian-speaking agent located in an Italian tourism agency, and an English-, German- or French-speaking customer at an arbitrary location. The third version (showcase 2b) was developed to evaluate the portability of the Nespole! STST system to new domains, and targeted first medical-aid assistance.

THE NESPOLE! DATA COLLECTION
For development and evaluation purposes, 201 speakers acted as customers or agents in a monolingual setting, enabling the collection of 306 dialogues that were used to train the linguistic modules of the three Nespole showcases. Another 77 subjects were involved in the two user studies described below, which tested the effectiveness of the multimodality-multilinguality combination, and assessed the communicative features of the systems. More details can be found in the NESPOLE! web site [4].

FIRST USER STUDY
The goal of the first user study (US1) was to evaluate the added value of multi-modal input in the multilingual NESPOLE! system [2].

During pre-tests preceding US1, 25 users were recorded in 21 monolingual and 29 multilingual dialogues using a first version of Showcase1. The analysis of user behavior and users' comments allowed us to improve the interface and the HLT modules of the system: Zoom and scroll functions were added, gestures drawn on maps by subjects could be saved and feedback for users was enhanced. The resulting system (Showcase1) was frozen and used for US1.

Method
14 German-speaking and 14 English-speaking novice users interacted with seven Italian-speaking travel agents in a push-to-talk mode, producing 28 dialogues. We compared two conditions: in the first condition the users could utilize the multimodal (MM) facilities (gesture drawing); in the second condition they had to rely only on speech (speech-only (SO)). All 35 users were given a task description and completed a questionnaire after their session.

SECOND USER STUDY
The second user study (US2) analyzed features and conflicts in multilingual and multimodal conversation, focusing on the significance of gestures. US1 had already shown clear advantages of MM over SO, and there were no demonstrated differences due to languages. Therefore, we limited our attention to the English-Italian pair, and compared the results of US1 and US2.

Method

Seven English customers interacted with three Italian agents using the fully implemented Showcase2 system in the MM condition, resulting in seven recorded dialogues. The scenario differs from US1 in providing an improved interface. The given budget became "budget per vacation package" while in US1 it was "budget per double room".

RESULTS

The comparison between the results of US1 and US2 shows improvements in terms of dialogue success and interface usability.

US1 showed that MM reduces the necessity for speakers to repeat turns, especially when the speakers were talking about spatial information (20% of repetition of spatial information turns in SO vs. 15% in MM). The improved interface of Showcase2 reduced repeated turns with spatial information in US2 to 13%. Moreover, in US1 a turn was repeated two times on average and only 1.6 times in US2.

Non-successfully translated turns with spatial information were at 31% in US1 for SO, and at 19% for MM. The latter figure was confirmed in US2.

Misunderstandings whereby speakers talked at cross-purposes along a couple of turns occurred in seven of the SO dialogues of the US1 dialogues. Two MM dialogues contained such misunderstandings, while the US2 dialogues showed none.

Gestures were used every 10 turns in US1, and every 8 turns in US2 on average. Gesture timing clearly shifted: In US1, all gestures were deployed after a given spoken turn was finished; in US2, they occurred before, during and after corresponding turns. The higher frequency of gestures, the improved integration and the newly achieved simultaneity of gestures in a spoken dialogue turn show that gesture deployment for MM became more natural in US2 than in US1.

US2 dialogues are shorter and contain fewer turns (US1: 35 min, 74 turns; US2: 23 min, 53 turns in average). Although a single turn contains more word tokens in US2 (8.2 vs. 5.8 in US1), we found a tendency towards more turns per minute in US2 (2.3) than in US1 (2.1), which is due to the longer waiting time for transfers and translations in US1.

CONCLUSION

Users influenced the development of the NESPOLE! system directly through comments, suggestions and questionnaires, and indirectly through the results of the user studies and analysis of observed problems during data collection:

The improvement of map saving, map transfer and the reduction of waiting time was mainly due to direct suggestions and comments of users during US1 and the monolingual data collection. The implementation of a new mechanism of map transfer dramatically reduced the time needed to visualize maps.

It was also observed during data collection sessions, that users had difficulties in understanding the feedback on the recognition of their contributions provided by the interface. They also had to blindly wait for a considerably long time not knowing if their contribution was still transferred, already successfully translated or whether the partner was also sending a dialogue contribution in parallel. This lead to mistreatments of the system and overlapping transfers resulting in system errors. Dialogues consisted mainly of pure waiting time where users got bored and distracted. Improving the NESPOLE! user interface by changing labels and positions of feedback windows, adding a web cam video and implementing progress bars helped to overcome errors due to incorrect usage of the system. Users were more involved and even if there are still waiting periods, users stay focused and busy with observing status and translation processes.

The results of the user studies show that the deployment of multimodality supports human interaction and translation in cross-lingual conversations better than uni-modal conditions. These advantages are increased as natural interaction with the system improves.

ACKNOWLEDGMENTS

Additional authors are Loredana Taddei, AETHRA S.r.l., Italy, and Walter Gerbino, Dept. of Psychology, University of Trieste, Italy. Special thanks to all the other NESPOLE! researchers and subjects involved in data collection, system development and user studies and Robert Isenberg for copy editing.

The work described in this paper has been partially supported by National Science Foundation under Grant number 9982227, and by the European Union under Contract number 1999-11562 as part of the joint EU/NSF MLIAM research initiative. Any opinion, suggestion and recommendation expressed in this paper are those of the authors and do not necessarily reflect the views of the EU or of the NSF.

REFERENCES

[1] A. Lavie et al. (2001) "Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Application", in Proceedings of HLT '01 (San Diego, U.S.A. , 2001).

[2] E. Costantini, F. Pianesi, S. Burger, (2002, II). "The Added Value of Multimodality in the NESPOLE! Speech-to-Speech Translation System: an Experimental Study". In proceedings of ICMI '02 (Pittsburgh, U.S.A., 2002).

[3] L. Taddei, E. Costantini, A. Lavie, (2002). "The NESPOLE! Multimodal Interface for Cross-lingual Communication - Experience and Lessons Learned". In proceedings of ICMI '02 (Pittsburgh, U.S.A., 2002)

[4] http://nespole.itc.it