Multimodality in a speech to speech translation system.

Preliminary results of an experimental study

Susan Burger (Carnegie Mellon University)

Erica Costantini (University of Trieste)

Walter Gerbino (University of Trieste)

Fabio Pianesi (ITC-irst)

Overview

- The NESPOLE! Project
  - Project's objectives
  - NESPOLE!'s infrastructure
  - HLT modules and multimodality
  - IF
- The study
  - Scenario and experimental design
  - Analisys of the data
  - Conclusions

Introduction

The project

- **NESPOLE!** is co-financed by the European Union and the National Science Foundation within the 5th Framework Programme.
- It started in February 2000 and will end in December 2002.
- NESPOLE!'s partners are: ITC-irst; Carnegie Mellon University – Language Technologies Institute; University of Karlsruhe – Interactive System Labs; Université Joseph Fourier (Grenoble); AETHRA (Ancona); APT (Trento)
- **NESPOLE!**'s main purpose is to show the feasibility of multilingual (through spoken language translation) and multimodal communication in the context of future services in the field of e-commerce and e-service.

Project's objectives

General

- NESPOLE! aims at providing a system capable of supporting **advanced needs in e-commerce and e-service** by resorting to automatic speech-to-speech translation and multimodal interaction.
- NESPOLE! does not only address **accuracy of translation,** but extends also the **ability of two humans to communicate ideas, concepts, thoughts and to jointly solve problems.**
- NESPOLE! will also provide for **non-verbal communication** by way of multimedia presentations, shared collaborative spaces and multimodal interaction and manipulation of objects.

Introduction

The workplan

Two major sets of activities spanning the whole temporal extent of the project:

- Study, development and evaluation of HLT modules (speech recognition, intermediate representation construction, sentence generation and syntesis)

- Activities related to multimedia/multimodality issues, and its impact on speech-to-speech translation settings.

Project's objectives

Scientific

ROBUSTNESS: capability of dealing with spontaneous speech and incomplete information.

SCALABILITY: in the same domain (Tourism).

CROSS-DOMAIN PORTABILITY: from Tourism domain to Help-desk.

MULITIMODALITY: exploring the use of multimodality in a multilingual human-to-human communication setting.

Two showcases

Showcase1 addresses a travel scenario, supporting the interaction, through the web, between a client and a destination agent.

Showcase2 is currently being defined. Most probably: conversation between a patient and a doctor.

Methods and technical overview

Infrastructure

Support for geographically distributed Language Specific HLT Servers, customers and agents.

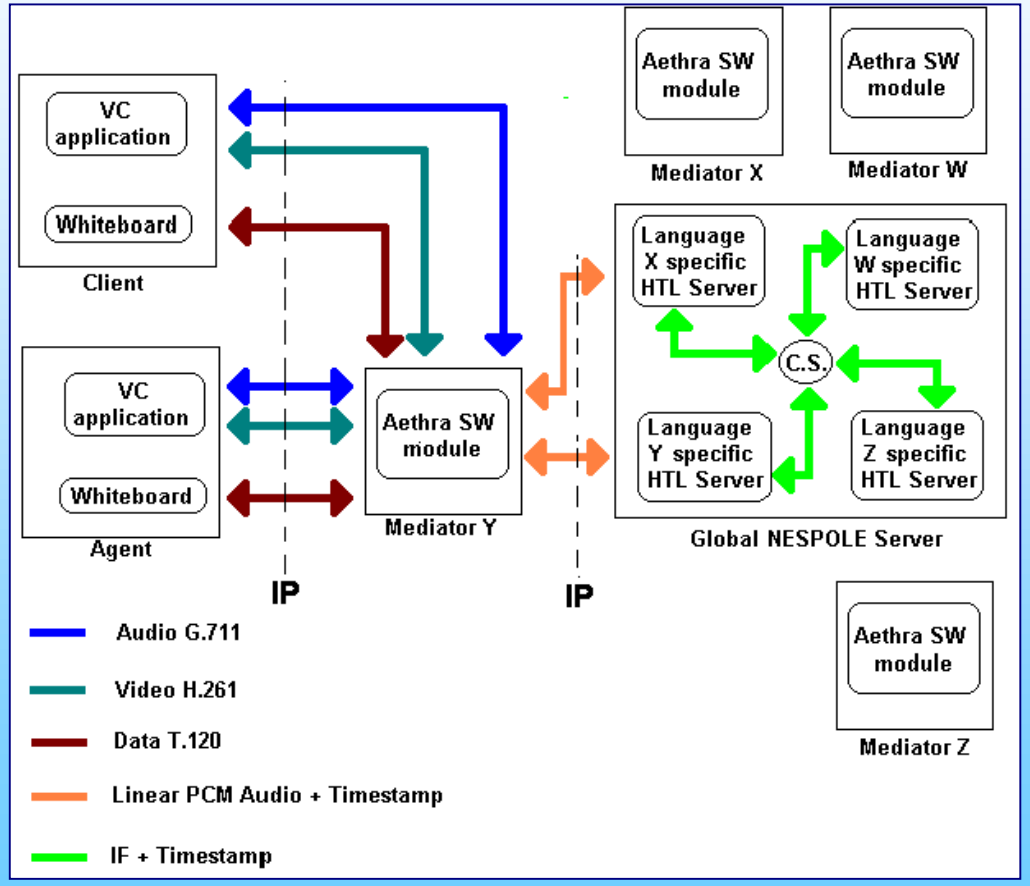
Complete structural symmetry between Agents and Customers.

Thin clients.

Monitoring tasks distributed among four distinct hosts.

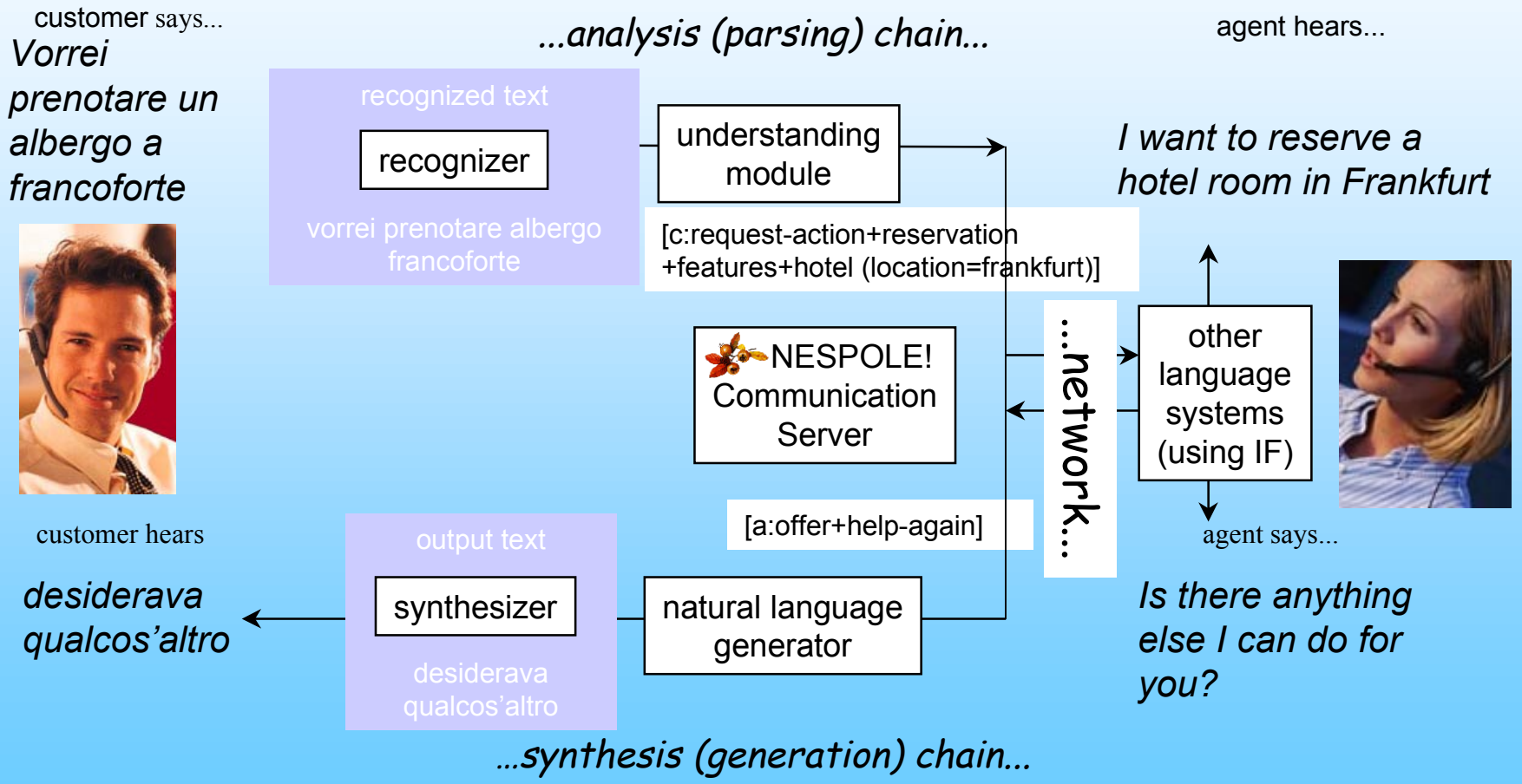
Methods and technical overview

The architecture of NESPOLE!



Methods and technical overview

The HLT Servers' architecture



Methods and technical overview

HLT modules

The overall philosophy of the project is to leave each partner free to develop the modules for its own language according to its preferences.

The only constraint, is that the basic issues of robustness, support for scalability, and portability across domain be addressed.

This gives the consortium the possibility of experimenting with, and comparing, a **range of approaches to speech and language analysis and language generation.**

IF
Intermediate Representation Formalism

A lot of work

Goals pursued:

- a general-purpose IRF to be used in conjunction with a more domain-oriented interlingua.
  the generic part exploits a frame-like representation. WordNet 1.6 provides the conceptual repertory.
  Important: the interplay between the general-purpose and the domain-oriented IRF.
- updates and improvements to the domain-oriented IF developed within CSTAR-II, to cope with the new requirements of NESPOLE!.
  - Extension of coverage to the new features of the application scenarios;
  - improvements over existing encoding for such linguistic information as referents novelty, numbers, nominals.

Scenario

The scenario for the first showcase involves an agent (Italian speaker), and a client (English, German or French speaker).

CLIENT

AGENT

Scenario

Showcase 1 is concerned with “Winter Accommodation in Val di Fiemme”.

- *winter accommodation for skiers is one of the typical tourist task for Trentino;*
- ***accommodation** is a field for which every partner has many acoustic and linguistic data;*
- *the scenario provides for rich interaction on many topics (e.g., local directions, location of ski rentals and parking; hotel facilities, children entertainment and menu), etc.*

Scenario

The considered scenario also offers good grounds for **experiments with multimodality**, being suitable to the use of

- pictures,
- videos,
- web pages

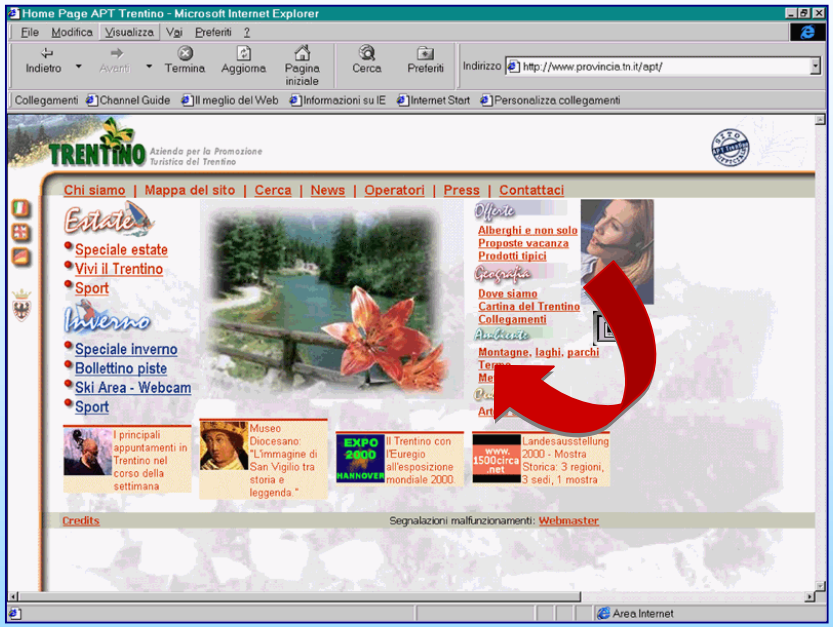
to describe places, and of

- gestures and drawings

to give directions.

Scenario
CLIENT screen

- The customer wants to organise a trip in Trentino.
- She starts by browsing APT web pages to get information.



- When the customer wants to know more about a particular topic or prefers to have a more direct contact, the speech-to-speech translation service allows her to interact in her own language with an APT agent.
- A videoconferencing session can be opened by clicking a button
- The dialog starts.

Scenario

- Both customer and agent have thin clients (with whiteboard)
- The customer's terminal connects to the Italian (Agent side) mediator, which acts as a multimedial dispatcher.
- The mediator
  - opens a connection with APT agent
  - transmits web pages
  - sends the audio to the appropriate HLT servers.
  - buffers and transmits gestures from the client to the agent and vice versa.
- Feedback facilities provide full control by both parties on the evolution of the communicative exchange.

Multimodality

- Gestures are performed by means of a tablet and/or a mouse on maps displayed through the system's whiteboard.
- Anchoring between gestures and language is obtained through a simple 'time-based' procedure.

More complex procedures, aiming at 'conceptual' anchoring have a greater impact on HLT modules. Their investigation has been postponed.

Multimodality
Usability study

- Goal: the impact of multimodality in a ‘real’ speech-to-speech translation environment

- Evaluation of the added value of multimodality in a multilingual and multimedial environment.

Multimodality
Previous results

- The advantages of multimodal input over speech-only input includes faster task completion, fewer errors, fewer spontaneous disfluences, strong preference for multimodal interaction (Oviatt, 97)

- when combined with spoken input, pen-based input can disambiguate badly understood sentences (Oviatt, 2000)

Comparison between the performances of two versions of the system:

- SO (Speech Only) version: multimedia with spoken input.
- MM (Multimodal) version: multimedia with spoken and pen-based input.

Multimodality - experiment

Hypotheses

- Pen-based input increases the probability of successful interaction, reducing the impact of translation errors
- The advantages of multimodal input are more relevant when spatial information is to be conveyed.
- The greater complexity of the the MM system does not prevent users from enjoying the interaction (and from evaluating it friendlier and more usable than SO system)

Scenario

Winter holydays in Val di Fiemme

A German or American speaker connects to the Trentino tourist office board (Italy) to ask for information about, and plan his/her holiday in Val di Fiemme

Experimental Design

MODALITY x LANGUAGE

MODALITY:

- SO (Speech only)
- MM (Multimodal)

LANGUAGE:

- English
- German

Experimental Design

Users: Customers

- TOTAL NUMBER: 14
- FEATURES:
  - English and German speakers
  - similar level of computer literacy and web expertise
  - paid volunteers
- DESIGN: between (each client took part in one dialogue and experienced only one modality)
- Sex: balanced across conditions

Table 1. Group composition

	E	G	sex
MM condition	4	3	F
	3	4	M
SO condition	3	2	F
	4	5	M
Sum	14	14	

E = English speakers; G = German speakers

Experimental Design

Users: Agents

- TOTAL NUMBER: 7
- Italian volunteers (not involved in the Nespole! Project) acting as Trentino tourist board agents
- DESIGN: within (each agent took part in more than one dialogue, and experienced both modalities)
- Sex: balanced across conditions and languages

Experimental Design

Dependent Variables

Variables targeted

- spoken input
- gestures
- effectiveness of the dialogue*
- usability self-reports

* Only for English dialogues

Speech

Spontaneous events:

- A-grammatical phrases (repetitions, corrections, false starts)
- empty pauses (silence, breathing)
- filled pauses (vowels, nasal, other)
- human noises (laugh, noise)
- word interruptions (speaker)
- understandability
- technical breaks (word break, word missing)
- turn breaks (the utterance is broken)

Speech

TURNS AND WORDS

- turns per dialogue
- tokens (spoken words) per dialogue
- types (vocabulary) per dialogue
- tokens per turn
- types per turn
- token/type rate (how many words were used before a new word was introduced)
- returns to topics already treated

Pen-based Gestures

Number and types of collected gestures:

- loading of an image*
- scroll*
- zoom*
- running a browser*
- selection of an area (only MM condition)
- pointing to an area (only MM condition)
- connecting different areas (only MM condition)

* in SO modality too: they are not properly multimodal inputs, but commands concerning multimedia

Dialogue effectiveness*

- number of successful turns
- ambiguities concerning place names (ski-areas, towns, hotels)
- reached goal: did the client find the hotel which meets his/her expense budget?

* Only for English dialogues

Usability self-reports

- S.U.S. (System Usability Scale) (agents and clients)
- Preference concerning experimental conditions (agents)

Experimental Design

Pre-tests

- MONOLINGUAL (ITA to ITA; ENG to ENG)

  Goal: collection of multimodal dialogues (n=20) for system development and for defining the task

- TECHNICAL TESTS

  Goal: testing architecture and connection

- INTERLINGUAL (ENG to ITA; GER to ITA)

  Goal: testing of language coverage; adjustments of task and instructions; testing of recordings tools; agents training

Experimental Design

Instructions for customers

Customers received written instructions concerning:

- goals of the experiment;
- information about how the system works;
- description of: interface, allowed inputs, system feed-backs, microphone managing;
- advises about most frequent system problems;
- in case of MM condition, clients were invited to use the pen for about 5 minutes before starting

Experimental Design

Task

The customer is asked to imagine being in the following situation:

- It is the end of November. You are going to spend a holiday in Val di Fiemme with a friend. Val di Fiemme is a region in northern Italy where you can find several ski aereas and resorts (villages).
- You are planning to go during the second week of December.
- You wish to go alpine-skiing and ice-skating.
- You would like to sleep in a three-star-hotel for 7 nights.
- You want to have half board accommodation (bed-and-breakfast and dinner)
- You are planning to go during the second week of December.
- Your available budget is at about 200.000 Italian Lire per night for the hotel *(this is about 90 US dollar)*.
  You want a double room.
- You will reach Val di Fiemme by airplane and bus. You already know about flight connections and bus transfer to Val di Fiemme.
- In Val di Fiemme, you plan to use public transportation.

The customer's task is to ask the Trentino tourist board office for more information and to choose:

- a TOWN with an ice-skating facility, and close to a ski-area
- a HOTEL close to a bus stop or a ski area. The hotel should meet the described requirements and the available budget

The client writes down a list of questions he/she would like to ask to the agent

Experimental Design

Instructions for agents

- TRAINING: agents were trained during the pre-tests.
- INFORMATION: agents received a table with information concerning two different towns and three hotels for each town (the presentation order of the options was balanced among conditions; all hotels, except one, are out-of-budget)
- agents were asked to give only information expressly asked by customers

Experimental Design

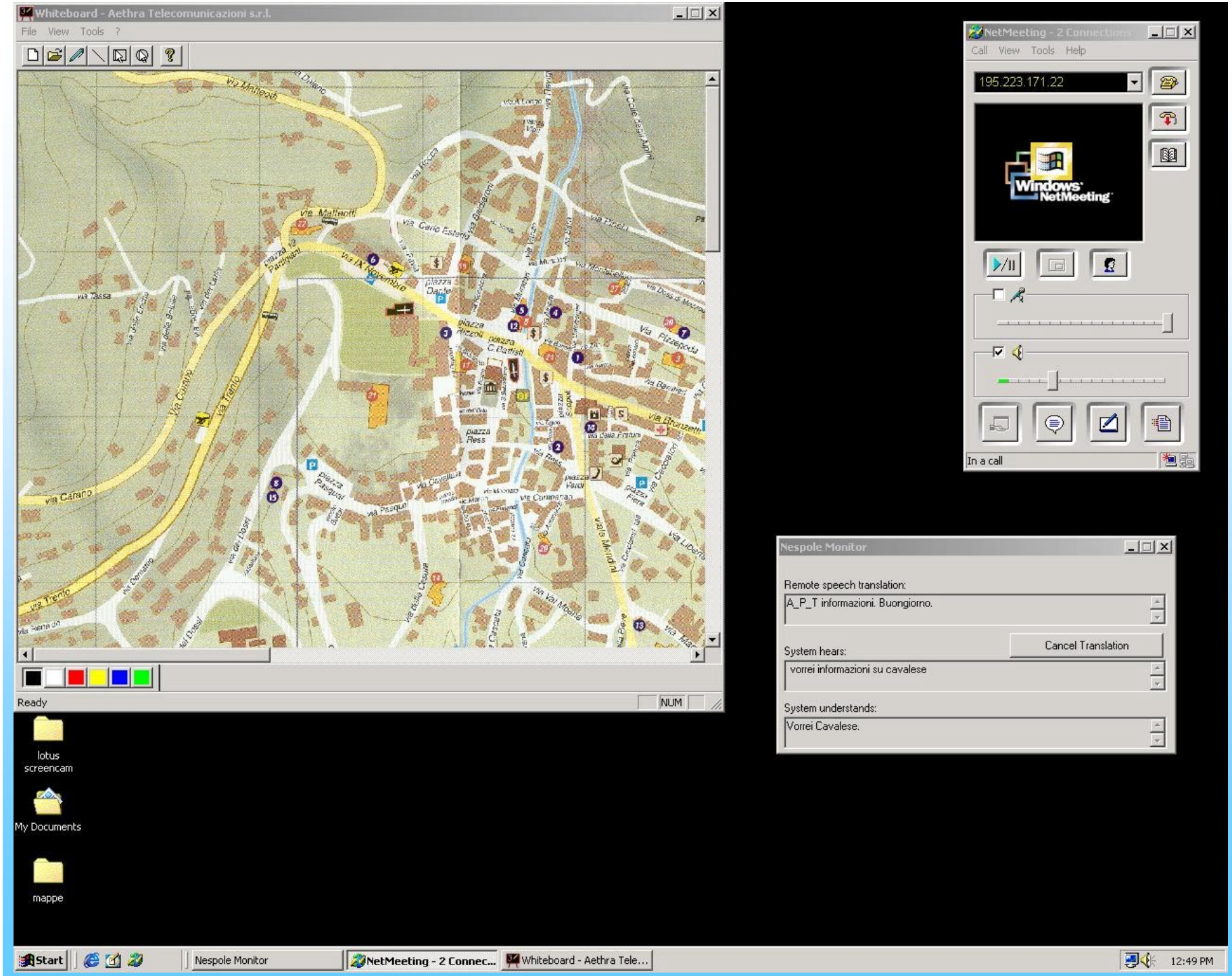
Material

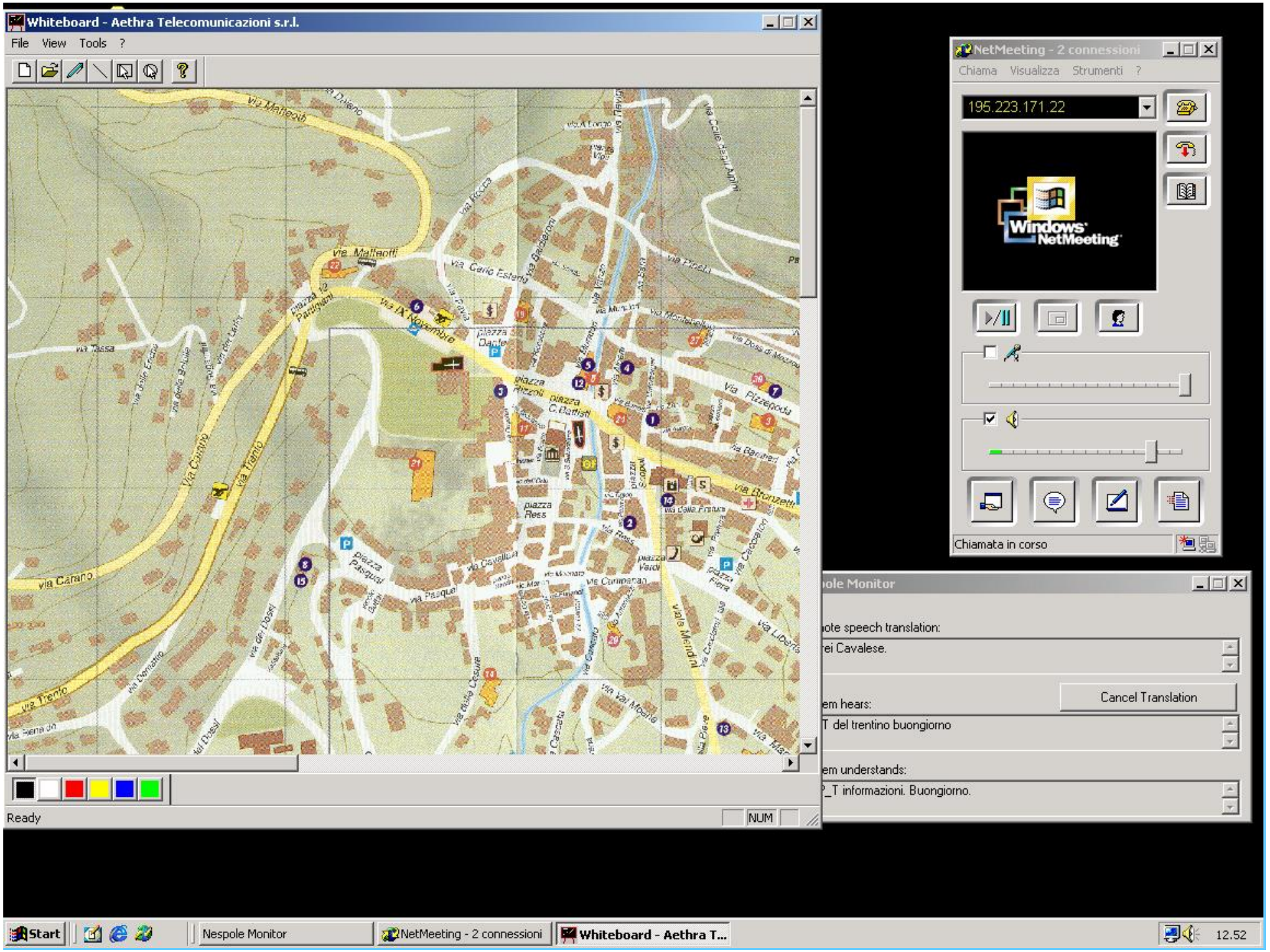
- Microphone
- Pen and tablet
- 3 maps
- Two web pages
- Same translation systems for the two conditions
- Different instructions for agents and customers

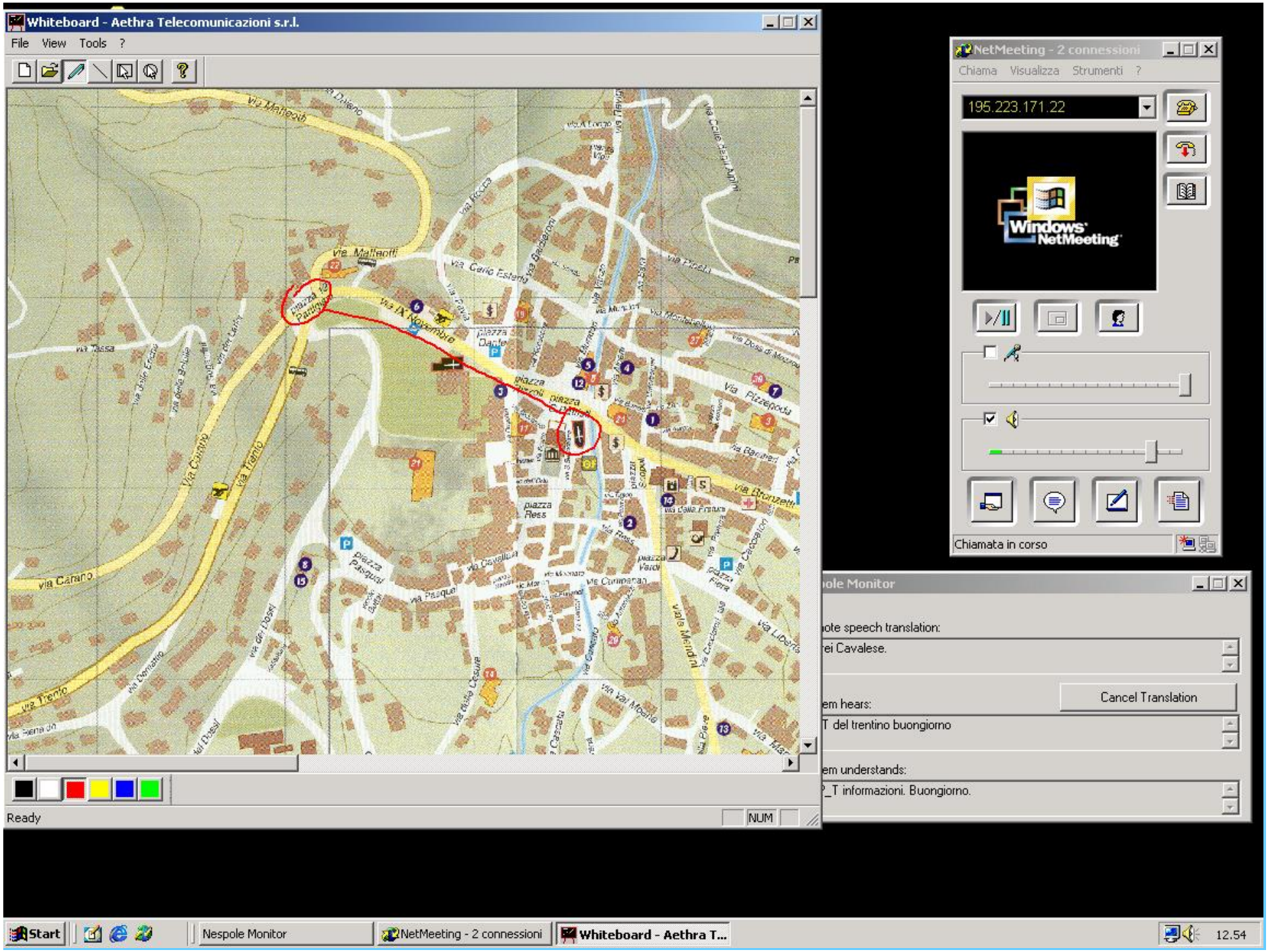
Experimental Design

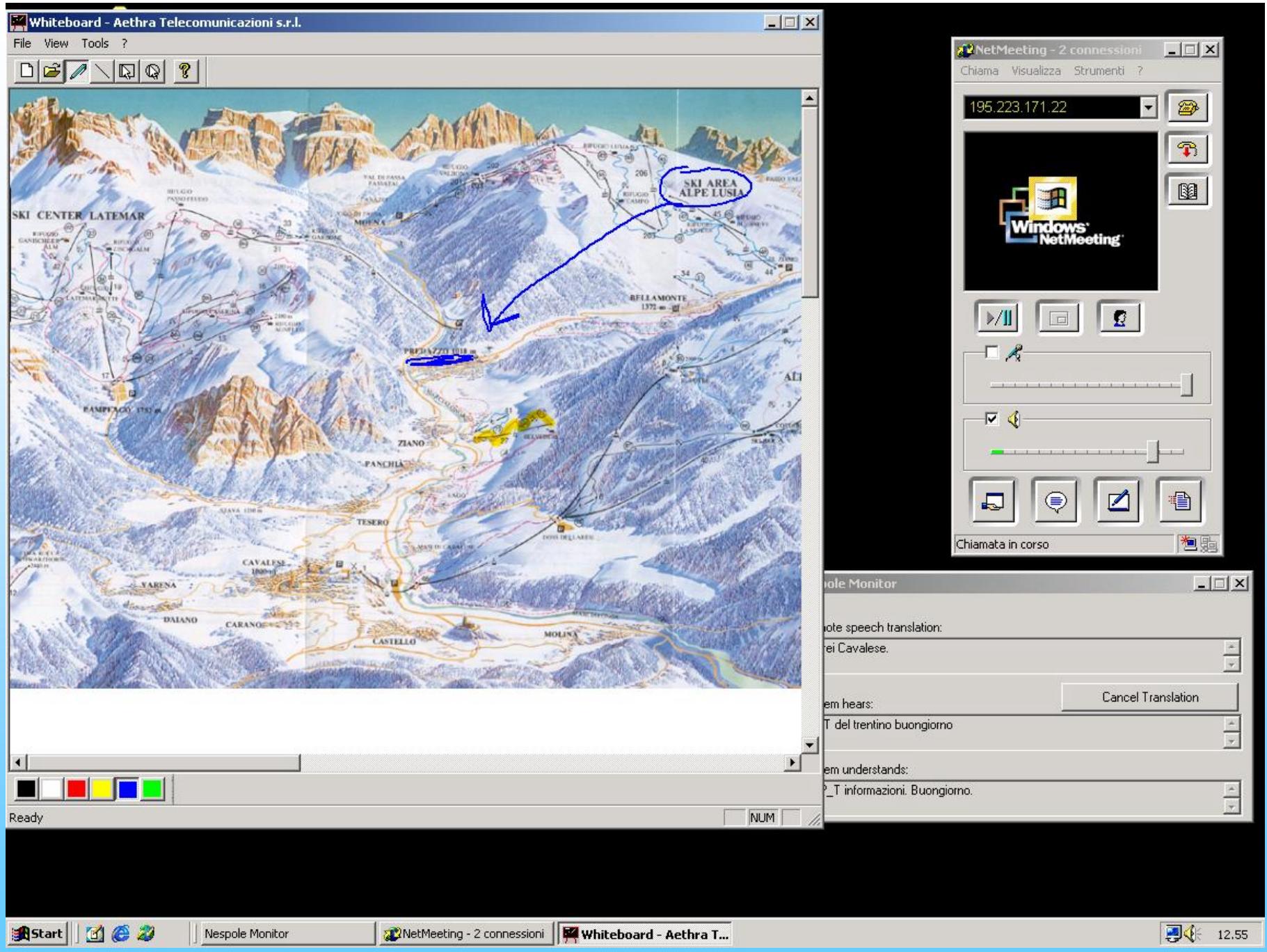
Material -screen

- Netmeeting window with
  - Push-to-talk button
  - Check-uncheck button
- Feedback window with
  - Hypothesed string
  - Hypothesed meaning
  - Textual translation of remote speech


Whiteboard - Aethra Telecomunicazioni s.r.l.
File View Tools ?
Ready
NUM
NetMeeting - 2 Connections
Call View Tools Help
195.223.171.22
Windows NetMeeting
In a call
Nespole Monitor
Remote speech translation:
A_P_T informazioni. Buongiorno.
Cancel Translation
System hears:
vorrei informazioni su cavalese
System understands:
Vorrei Cavalese.
lotus screencam
My Documents
mappe
Start
Nespole Monitor
NetMeeting - 2 Connec...
Whiteboard - Aethra Tele...
12:49 PM


Whiteboard - Aethra Telecomunicazioni s.r.l.
File View Tools ?
Ready
NUM
NetMeeting - 2 connessioni
Chiama Visualizza Strumenti ?
195.223.171.22
Windows NetMeeting
Chiamata in corso
ole Monitor
ote speech translation:
ei Cavalese.
Cancel Translation
em hears:
T del trentino buongiorno
em understands:
_T informazioni. Buongiorno.
Start
Nespole Monitor
NetMeeting - 2 connessioni
Whiteboard - Aethra T...
12.52


Whiteboard - Aethra Telecomunicazioni s.r.l.
File View Tools ?
Ready
NUM
NetMeeting - 2 connessioni
Chiama Visualizza Strumenti ?
195.223.171.22
Windows NetMeeting
Chiamata in corso
ole Monitor
ote speech translation:
ei Cavalese.
em hears:
Cancel Translation
T del trentino buongiorno
em understands:
_T informazioni. Buongiorno.
Start
Nespole Monitor
NetMeeting - 2 connessioni
Whiteboard - Aethra T...
12.54


Whiteboard - Aethra Telecomunicazioni s.r.l.
File View Tools ?
SKI CENTER LATEMAR
SKI AREA ALPE LUSIA
PANCHIÀ
ZIANO
TESERO
CAVALESE
VARENA
DAIANO
CARANO
CASTELLO
MOLINA
Ready
NUM
NetMeeting - 2 connessioni
Chiama Visualizza Strumenti ?
195.223.171.22
Windows NetMeeting
Chiamata in corso
Nespole Monitor
ote speech translation:
ei Cavalese.
Cancel Translation
em hears:
T del trentino buongiorno
em understands:
_T informazioni. Buongiorno.
Start
Nespole Monitor
NetMeeting - 2 connessioni
Whiteboard - Aethra T...
12.55

Experiment - results

Successful dialogues

CANCELED DIALOGUES: N = 22

- client didn't show up: 3
- interrupted (connection or hlt servers crashes): 4
- connection problems (connection failed): 4
- the system was not yet 'frozen': 5
- incomplete recordings: 6

FULLY RECORDED DIALOGUES: n = 28

- delays due to connection problems (about 20 minutes): 3
- interruption and restart during dialogue: 3
- synthesis crashed 10 minutes before the end of the dialogue(but dialogue contined in 'text' mode): 2

Experiment - results

Speech-related variables

- No significant differences among conditions as to spontaneous events, turns and words figures, dialogue lenght.
- One spoken turn every 33 seconds (average) in both conditions.
- Average duration per dialogue:

SO=38 min. MM=28 min t-test=0.12

Experiment - results

Speech-related variables

- Customer: vocabulary per dialogue.

  German 103, English 81.86 t-test=0.037

- Customer: words per dialogue

  Male: 260, Female 205, t-test=0.033

- Customer: vocabulary per dialogue:

  Male: 100.3, Female=82, t-test=0.05

- Agents: words per turns,

  English clients=7.71, German clients=6.5, t-test=0.05

Experiment - results

Successful turns
Only English

- Real turns (excluding non-understandable case)

  SO: 486 (83%) MM: 368 (79%)

- Average duration of real turns

  SO: 33,78 secs MM: 32,45 secs

- Number of repeated turns (both immediate and later repetitions):

  SO: 79 MM: 59 (t-test 0.09)

Experiment - results

Speech-related variables

	Num. of topics	Turns per topic	Returns number	Returns per topic	Return rate*	n. Spatial returns
SO	66	8.65	30	0.46	19	14
MM	56	8.36	15	0.27	31.2	5
T-test					0.086	0.07

* Return rate= number of turns / number of returns

Experiment - results

Gesture-related variables

- All gestures (but 2), performed by agents
- Total gestures:

  SO: 63 MM: 182

- Few or no deictics used. Mostly accompanying speech (*I'll show it to you on the map*)

Experiment - results

Gesture-related variables

Average figures for gestures:

- loading of an image: 2,7 (MM and SO. No significant differences)
- scroll: 1,7 (both MM and SO. No significant differences)
- zoom: 0
- running a browser: 0,4 (both MM and SO, No significant differences)
- MM-only gestures: 7
  - selection of an area: 4,71
  - pointing on an area: 1,36
  - gestures connecting different areas: 1

Experiment - results

Ambiguities

Number of dialogues containing ambiguities concerning place names (ski-areas, towns, hotels)

	MM	SO
yes	2	5
no	5	2

Experiment - results

Goals achievement

No differences in the number of dialogues in which the client found/didn't find the hotel meeting the requirements

	MM	SO
yes	2	2
no	5	5

Experiment - results

Usability

- No differences among conditions as to S.U.S.* scores.
- No differences between clients group and agents group as to S.U.S. scores.
- Average score: 55 **

* System Usability Scale (developed by Digital Equipment Co. Ltd, Reading, UK)

** S.U.S. scores have a range of 0 to 100

Experiment - results

Usability

- Strong preference of agents for multimodal interaction
- Weak preference of agents for the English Language

AGENT	pref SO	pref. MM	pref ENG	pref Ger
1		**X**	x	
2		**X**	x	
3		**X**	**X**	
4		**X**	**X**	
5*				
6*				
7*		**X**		

X = strong preference
x = weak preference
* Agents n.5, 6, 7 took part in 3 or 4 dialogues (less than half respect to the other agents); n. 5 and 6 have not preferences; n. 7 has not preference concerning language)

Experiment

Provisional conclusions

- The presence/absence of multimodality does not seem to systematically affect the dependent variables
- MM seems to have some effect on speech-related variables, though this is rarely statistically significant.

- Tendency for dialogues to be shorter in MM than in SO
- Tendency for repeated turns to be fewer in MM than in SO
- If returns can be taken as an indicator of dialogue fluency, then there is a tendency for fluency to be better in MM than in SO.
- Moreover, this is even clearer for dialogue segments dealing with spatial information.

Experiment

Provisional conclusions

- No, or very rare, spontaneous use of deictics.
- All MM gestures have been used by agents, with a clear preference for area selection.
- Tendency for MM to exhibit less ambiguity
- Moreover, when present, the ambiguity was immediately solved by resorting to MM resources.
- However, there doesn't seem to be a difference in effectiveness (goal achievement) between SO and MM.
- Strong preference for MM by agents.

Experiment

Provisional conclusions

- There aren't yet systematic data about interactions between conditions.
- In many cases, data about German are still missing.
- Not a highly structured task.
- Great variability due to the 'reality' of the experimental setting (high variance)
- Perhaps, more subjects could balance this, and disambiguate borderline cases.

Experiment

Provisional conclusions

- Pen-based input increases the probability of successful interaction, reducing the impact of translation errors

- The advantages of multimodal input are more relevant when spatial information is to be conveyed.

- The greater complexity of the the MM system does not prevent users from enjoying the interaction (and from evaluating it friendlier and more usable than SO system)

NESPOLE! Will be at the next IST Conference in Düsseldorf.

If you come, please, visit us and play with the system!!