The Added Value of Multimodality in the NESPOLE! Speech-to-Speech Translation System: an Experimental Study

Erica Costantini
Department of Psychology
University of Trieste, Italy
costanti@psico.units.it

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

Susanne Burger
Interactive Systems Laboratories
CMU, Pittsburgh, USA
sburger@cs.cmu.edu

Abstract

Multimodal interfaces, which combine two or more input modes (speech, pen, touch…), are expected to be more efficient, natural and usable than single-input interfaces. However, the advantage of multimodal input has only been ascertained in highly controlled experimental conditions [4, 5, 6]; in particular, we lack data about what happens with 'real' human-human, multilingual communication systems. In this work we discuss the results of an experiment aiming to evaluate the added value of multimodality in a "true" speech-to-speech translation system, the NESPOLE! system, which provides for multilingual and multimodal communication in the tourism domain, allowing users to interact through the internet sharing maps, web-pages and pen-based gestures. We compared two experimental conditions differing as to whether multimodal resources were available: a speech-only condition (SO), and a multimodal condition (MM). Most of the data show tendencies for MM to be better than SO.

1. Introduction

Previous research using Wizard-of-Oz technique demonstrated that, when interacting on spatial tasks, the performances of users sensibly improve if multimodal input is available, leading to faster task completion, fewer input disfluences, less complex language and greater satisfaction [4]. Moreover, it was found that multimodal interaction occurs more frequently in case of spatial location commands [5]. It is also suggested that well-designed multimodal systems can integrate complementary modalities in a way that supports significant levels of mutual disambiguation and recovery from errors [6].

These results were obtained in highly controlled experimental conditions, in a monolingual setting. Data are still lacking that concern the extent to which those result can be replicated in scenarios relying on "real" systems for human-human multilingual communication. In particular, it is important to know how robust the mentioned improvements are vis-à-vis disturbing factors such as system's failures, time lag due to network traffic, etc. At the same time, when multilinguality is realized through speech-to-speech translation (STST), it is crucial to ascertain whether the use of pen-based gestures can help to overcome the weaknesses of the underlying Human-Language-Technologies, providing synergies that the user can exploit to improve the quality and success of the interaction. We designed and executed an experiment, aiming to test:

- whether multimodality increases the probability of successful interaction, even with prototypes of 'real' multilingual systems, when spatial information is the focus of the communicative exchange;
- whether multimodality supports a faster recovery from recognition and translation errors.

The 'real' system we exploited is the first showcase of NESPOLE! (NEgotiating through SPOken Language in E-commerce), a jointly EU/NSF funded project dealing with STST in e-commerce and e-services [1, 2, 3]. The languages addressed are Italian, German, English and French. The scenario involves an Italian-speaking agent located in an Italian tourism agency (APT), and an English-, German- or French-speaking customer at an arbitrary location. The two communicate through the Internet using thin terminals (PCs with sound and video cards and H323 video-conferencing software), and can share web pages and maps by means of a special White Board. NESPOLE! provides for multimodal communication, allowing users to perform gestures on displayed maps, by means of a tablet and a pen.

2. Method

2.1. Experimental design

Two experimental conditions were considered:

- a speech-only condition (SO), involving multilingual communication and the possibility for users to share images and maps through a White Board;

- a multi-modal condition (MM), where users could additionally perform pen-based gestures (pointing, area selection, connection between different areas) on shared maps to convey spatial information.

The experiment involved American English native speakers (located at Carnegie Mellon University, Pittsburgh) and German native speakers (located at University of Karlsruhe), who played the role of the customers. Both interacted with Italian native speakers (located at Irst, Trento), who were trained to act as tourist agents. This resulted in four experimental groups: German customer/SO, German customer/MM, English customer/SO, and English customer/MM.

2.2. Task and Instructions

The scenario of the experiment was modeled after one of the five different tourism scenarios, covered by Nespole! [7], enriched with spatial information. This scenario features a customer browsing the web pages of a tourist office in Trentino, Italy. When the customer wants more information, she clicks on a special button, which opens a direct, STST-mediated connection with a human agent. The customer's task in the experiment was to choose an appropriate location and a hotel within specified constraints, concerning the relevant geographical area, the available budget, etc. The agent's task was to provide the necessary information.

Customers received written information and instructions about the scenario, the task, the system's functionalities and interaction modalities. In the MM condition, we also demonstrated the White Board functionalities, and allowed users a few minutes to familiarize themselves with the optical pen.

Agents were given description cards with information about two locations in Val di Fiemme (a tourism resort in Trentino), and three hotels for each location. The agents received training and instruction in proper methods of response (kinds of answers allowed, style, etc.) so as to adhere as much as possible to what 'real' travel agents usually do. For the same reasons, only agents were allowed to send maps and web pages, as it is the tourism operator and not the customer who knows which resources can be helpful at which point, where they can be found, etc.

During the experiment, subjects wore a push-to-talk head-mounted microphone. Subjects could only hear the translated message of the other party.

NESPOLE!'s screen displayed three windows:

a) The Aethra® White Board window, set at 600x600 resolution, used to display maps. During the MM condition, the users were allowed to draw gestures on the shared maps using the White Board drawing functionalities, which include:
   - free-hand strokes to draw arrows, lines, circles, etc.;
   - lines to connect two point on the maps;
   - selection of areas, done by enclosing portions of maps within elliptical/rectangular shapes.

   It is additionally possible through the Aethra White Board to run a browser: as soon as the specific function is selected, a web-browser window is available on the screen, until the user closes it.

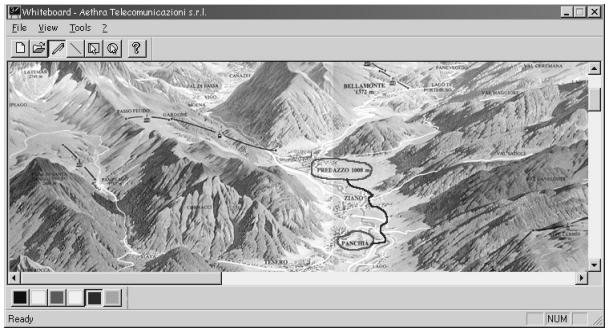

Figure 1. Aethra® White Board

b) The Feedback window, displaying useful information to the users concerning: the hypothesis string produced by the speech recognizer, as well as a string of information about the system's understanding of the speech.

c) The NetMeeting® window, allowing control over the usual features of this application. The window includes a push-to-talk button to activate or de-activate the microphone.

2.3. Participants

A group of 39 subjects participated in the experiment: 32 paid volunteers (16 American English and 16 German native speakers; sex balanced) played the role of the customer, and 7 native Italian speakers acted as tourist agents. All subjects had a comparable level of computer literacy and web expertise, as tested through by mean of a questionnaire.

2.4. Recordings, Transcriptions and Annotations

The collected data set consisted of 28 dialogues: 14 involving an American English customer and 14 involving a German customer; all dialogues involved an Italian agent. Each group consisted of seven SO and seven MM dialogues. The average duration of dialogues was 35 minutes (range: 19-59 minutes).

The audio files were transcribed in accordance to the VERBMOBIL conventions [7], using the TransEdit annotation tool. Besides orthographic words, transcription files contained:

- annotations for spontaneous phenomena: false starts/repetitions, empty pauses, filled pauses, human noises, word interruptions and breaks, turn breaks and

incomprehensible utterances. Technical interruptions are also marked;

- annotations for gestures (as three-line comments added at the end of the corresponding speech turn), including gesture identification, gesture description (based on the usage of White Board commands) and gesture goals (selection, pointing, connection, words). Gestures were manually annotated using videos recorded on the Italian side. For details concerning gesture annotation conventions see [8].

By comparing original and translated turns with their replies, we classified all turns into *successful, partially successful and non-successful*:

Successful turns had translations that were grammatically, syntactically and semantically accurate.

Partially successful turns had poor or bad translations, either because of grammatical or syntactical errors, or because some words were badly translated or not translated at all. The translation, however, managed to preserve enough of the original message to enable the targeted party to react properly.

A turn was labeled as **non-successful** if the other party could not understand any component of the original utterance.

Turn repetitions (where the speaker repeats her utterance because of errors made by the system) were counted as well.

3. Results: Language and Gestures

3.1. Language and Spontaneous Phenomena

We operationally defined a turn as a speaker's contribution between a switching-on and a switching-off of the microphone button in the NetMeeting® window of the Nespole! monitor. Word-tokens are occurrences of a given word-type — e.g., the sentences "Paul is the brother of John" and "John is the brother of Paul" contain 12 word-tokens and 6 word-types. The relevant digits for these variables are shown in Table 1.

Table 1: Average number of turns, tokens, types and tokens per turn, for each language

	Italian agent	German customer	English customer
Turns per dialogue	37	39	33
Tokens per dialogue	258	254	218
Types per dialogue	101	103	82
Tokens per turn	6.98	6.50	6.60

The average duration of dialogues was 35 minutes; dialogues did not significantly vary in duration, across either different modalities or languages.

We computed the percentage of spontaneous phenomena per word token. The resulting figures are very low: for seven of the eight annotated phenomena occurrences are always lower than 3% (mostly even below 1%). Only the percentage of empty pauses at the customer site is a bit higher, ranging from 6% to 10%. The low frequency of disfluences is probably connected to the push-to-talk procedure, which allows speakers enough time to plan their contributions, this way reducing errors and hesitations. We calculated a fluency score for each dialogue, as the weighted sum of the average frequencies for each class.

Average values and variance for dialogue length, turns, tokens, types, as well as for fluency score, are similar across agents and customers and across languages and modalities. ANOVA tests (p=0.05) run on the number of turns and on the fluency score (customers and agents separately) did not detect any effect of modality and/or language. Hence, we can conclude that language or modality had no significant effect on such linguistic features of dialogues as the number of turns and the fluency score.

3.2. Pen-Based Input

We counted the number of *selection, pointing* and *connection gestures* for each dialogue, and annotated which of the White Board functionalities (*free-hand, line* or *elliptical/rectangular selection*) was used. In addition we counted how often agents used the *free hand* modality to write words on the map, most of these being hotel or town names associated with selection or pointing gestures.

Table 2. Percentage of performed drawing gestures and used White Board functions (MM condition)

Drawings	% on all	*Mode*	*% on class*
Selection	61%	free-hand	65%
		elliptical	31%
		rectangular	4%
Pointing	19%	free-hand	100%
Connection	12%	free-hand	47%
		line	53%
Words	8%	free-hand	100%

The average number of *drawing gestures* per dialogue (MM condition) was 9. Given that the average number of turns per dialogue is 73, this means that gestures were performed on average every 8 turns. Such low ratios are probably due to the fact that interaction involving spatial information was confined to a few dialogue segments. Table 2 shows the distribution of each gesture.

The figures in the table do not distinguish between the agents' and the clients' contributions, given that the agents performed almost all the drawings (98,1%).

A clear preference emerges for area selections among the drawing gestures (61% of the total number of drawings), and for the free-hand mode.

3.3. Speech-gesture Association

Three classes of temporal integration patterns between gestures and speech were annotated: *immediately before*, *during* or *immediately after* the corresponding turn. Table 3 reports the relevant figures, for each class of gestures.

As can be seen, most of the gestures (79%) followed the speech turn, and none were performed during the turn. The typical sequence occurring when an agent wanted to use drawings (or to load maps or to send web pages), consisted of some kind of verbal anticipation of her intentions — e.g. "I'll show you the ice skating rink on the map" — followed by a switching off of the microphone, and then by gesture performance and feed-back request, for example, "Can you see the skating rink?". It can be argued that this particular sequence and the absence of gestures during the speech were influenced by the push-to-talk procedure and the time needed to transfer gestures across the Internet. More precisely, the verbal cues were meant to alert the other party that she had to wait for a forthcoming gesture, possibly refraining from speaking in the meanwhile. This procedure allowed agents enough time to perform the gesture and ask for feedback. In addition, both microphone on/off switching and drawing functions were performed by means of the pen device. It can be argued that managing both tasks nearly simultaneously further discouraged the simultaneous execution of speech and gestures.

Table 3. Percentages of gestures performed before, during and after the speech

Drawing gestures	Before	During	After
Selection	19%	0%	81%
Pointing	26%	0%	74%
Connection	20%	0%	80%
Word	33%	0%	67%
Sum drawings	21%	0%	79%

Few or no deictics were used. Sometimes the customer used indicator "here" to inform the agent that the map or the web page was on her screen ("the map is here"). No other relevant uses of deictics could be found. Agents preferred to resort to descriptive phrases that relied on visually available cues — e.g., "the skating rink is at the bottom right of the map", "I'm selecting it with the red color".

Those findings, too, seem related to the push-to-talk procedure. As already mentioned, users tend to avoid mixing gestures and speech. Thus, there was always a certain time lag between speech and gestures. Deictics, on the other hand, consist of linguistic markers (almost) concurrent with demonstrations (gestures). In the described situation, they would tend to be infelicitous, and rarely used.

4. Results: Dialogue Effectiveness

4.1. Goals Attainment

The goal of the customer is to book a hotel, meeting some assigned constraints (three-star hotel with half board accommodation, close to a bus stop or ski-area, no more that 108,5 Euro for a double room per night, etc). All available hotels were three-star hotels, and all prices included half-board accommodation. All possible hotels were out of the budget range, except for the two target hotels.

The number of successful dialogues was 24 (86%), without relevant differences among modalities. This demonstrates that the STST system is good enough for novice users to accomplish a task with minimal written instructions, very short initial training on the White Board, and no further assistance during the interaction. At the same time, multimodal communication did not provide any clear advantage on the completion of the particular task we chose.

4.2. Successful Turns and Turn Repetitions

We computed the percentages of each class of turns (see § 2.4) both on the total number of turns ("all turns") and on legal turns only. Legal turns are defined as turns discussing "legal" matters. A given topic was classified as illegal if it was not among those specified in the written instructions, even if it sounds reasonable within the given domain. For example our written instructions did not provide for questions about whether there is much snow in December, or whether anyone at the hotel speaks German, though these are *reasonable* questions in the tourism domain. Illegal questions were neglected to eliminate factors that could affect dialogue in unpredicted ways.

Finally, the same statistics were computed for the turns conveying spatial information ("spatial turns"). The expectation was that possible effects of MM on dialogues could be better demonstrated by focusing on turns containing spatial information.

Table 4. Percentages of successful, partially successful and non-successful turns on all turns, legal turns and spatial turns

	All turns	Legal turns	Spatial turns
Successful turns	29%	31%	29%
Partially successful turns	32%	35%	41%
Non-successful turns	39%	34%	29%

Table 4 reports average distribution for each class of turns across all turns, legal turns and spatial turns.

The percentage of non-successful turns for legal turns is slightly lower than that for all turns, which confirms our hypothesis that illegal topics have a misleading effect.

The same values decrease even more clearly when only spatial turns are considered, pointing towards a possible positive effect of MM on turn success. The decrease of unsuccessful turns within spatial segments, in fact, is associated with an increase of partially successful turns, but not of successful turns. This suggests that some factors could improve the communicative effect of otherwise poorly translated spatial turns, enabling the other party to react properly, and permitting to classify the relevant turn as partially successful rather than non-successful. The obvious candidates are gestures in the MM condition. This hypothesis is supported by table 5, which shows a tendency for MM to reduce the number of non-successful turns with respect to SO. This tendency is more evident in the case of spatial turns.

Table 5. Percentages of non-successful turns on all turns, legal turns and spatial turns, split across conditions

Percentages of non-successful turns	Eng. SO	Eng. MM	Ger. SO
All turns	33%	27%	34%
Legal turns	33%	26%	29%
Spatial turns	30%	19%	31%

Speakers often repeated turns in order to overcome system errors or misunderstandings. In our experiment, each repeated turn was repeated two times, on average. Table 6 reports the distribution of repeated turns according to the classification considered here.

As can be seen, repeated turns tend to diminish in the MM condition (11% vs. 17% for English, and 18% vs. 23% for German), when only spatial segments are considered. This is consistent with the conclusions above: MM increases the number of partially successful turns while decreasing the number of unsuccessful ones.

Table 6. Percentage of repeated turns on all turns, legal turns and spatial turns, for all groups

Percentages of repeated turns	Eng. SO	Eng. MM	Ger. SO
All turns	16%	16%	20%
Legal turns	15%	15%	20%
Spatial turns	17%	11%	23%

It is clearly possible to conclude that multimodality can increase the probability of successful interaction and support a better recovery from translation errors, as well as reduce the number of turn repetitions.

4.3. Dialogue Fluency

Speakers sometimes returned to previously discussed topics. When occurring frequently, those returns complicate the dialogue flow and decrease dialogue fluency. Returns are usually related to difficulties in successfully closing a dialogue segment. For instance, if the customer does not obtain clear answers to her questions, she may abandon the current topic and return to it later on, asking for further clarifications. Our hypothesis that MM positively affects dialogue fluency implies that it could help speakers in successfully close dialogue segments, thus reducing the need to reiterate old topics, and yielding fewer returns.

The average number of returns per dialogue is 3.6. We computed a *return rate* by dividing the number of turns by the number of returns. This rate indicates how many turns were spoken in average from one return to the next, and can be used as an index of dialogue fluency: the greater the index, the better the fluency. In the English dialogues the average return rate shows a clear tendency to be higher for the MM condition (31) than the SO condition (19). German dialogues show a very small difference between return rates in SO and MM conditions (respectively 21 and 24); though this differences is not as marked, it follows the same trend of that found in the English dialogues. We tentatively conclude, mainly on the basis of English data, that MM can indeed ameliorate dialogue fluency by reducing the number of returns.

4.3. Ambiguities

Sometimes during a dialogue agents and customers end up discussing different topics without being aware of that they are not talking about the same thing. Such a misunderstanding can last for many turns and may not even be clarified by the end of the dialogue.

Direct observations of agent/customer interactions suggested that MM (i.e. gestures on the whiteboard) could aid in the resolution of misunderstandings; to check this we counted the number of dialogues in which topic confusion occurred. The number of dialogues containing ambiguities concerning place names was higher in SO (7 dialogues, 50%) than in MM (3 dialogues, 21%). Thus, multimodality seems to be effective in preventing ambiguities, when compared with speech input alone.

Qualitative analysis of transcripts sharpens this point: transcripts reveal that some SO dialogues contain more than one ambiguity, which in many cases remained unsolved. In MM, the three dialogues with ambiguities contained only one of each, and those ambiguities were solved in a couple of turns: as soon as the agent felt that the customer had not properly understood, she availed herself of the MM functionalities to select and show the customer the place she was speaking about. In the same situation, under SO condition, the agent had to resort to language for clarification, this strategy being obviously affected by the limitations of the STST system. The frequent failures in this respect seem to show that the paraphrases or whatever used by the speaker to recover

from ambiguities were often outside the reach of the STST system. Hence, one of the main hypotheses of our study is further supported: multimodal input can indeed help overcome the limitations of STST systems.

5. Conclusions

Considering all the above-mentioned results, multimodal interaction seems not to affect the dialogue length, the number of spoken turns and words, and the number of disfluences and spontaneous phenomena. On the other hand, it seems quite capable of enhancing dialogue effectiveness. When spatial information is conveyed MM input is clearly better than SO in decreasing the number of ambiguities, repetitions and non-successful turns; in addition, it helps in solving misunderstandings and provides for a better dialogue fluency. Moreover, when explicitly asked to express a preference between the MM and the SO condition, the users who acted as agents (the only users who experienced booth interaction modality conditions) indicated a clear preference for the MM system version. However, this evidence in favor of the multimodal interaction is weaker than that established in previous researches. Most of our data show tendencies and progressions towards better results when using MM, the results do not have the force of significant statistical data.

The fact that we used a real system prototype instead of the Wizard-of-Oz is of primary importance to assess the results. The NESPOLE! system caused errors and failures during the interactions (often due to network traffic), which, in turn, resulted in abnormally high variance in the measured variables, thus lowering the power of the statistical tests. It is remarkable that, despite these adverse conditions, the task was completed in the great majority of cases, indicating the effectiveness of the system in supporting users, both in SO and in MM.

It should also be mentioned that we did not do more on the multimodal side - allowing users more room to use multimodal interaction, for instance - because the translation modules available at the time of the experiment would have not been capable of supporting this. The task we devised was the best compromise between the system's capabilities at that time, and the need to provide for true pen-based gestures.

The NESPOLE! consortium has already profited from these results, using observations and insights from the experiment to improve the system — e.g., by providing users with better support for feedback about the success of the different stages of the translation process. The dialogues and behavior of the subjects involved in the experiment deserve further investigation and will serve as a basis for additional improvements of the NESPOLE! and similar translation systems.

7. Acknowledgments

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.

The authors wish to acknowledge the contribution and support by the other participants in NESPOLE! to realization of the experiment described here. In particular, Alon Lavie, Chad Langley, Celine Morel, John McDonough, Florian Metze, Loredana Taddei, Roldano Cattoni, Francesca Guerzoni and Walter Gerbino.

8. References

[1] G. Lazzari, "Spoken Translation: Challenges and Opportunities", in *Proceedings of ICSLP' 2000*, Beijing, China.

[2] A. Lavie at al, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei and F. Balducci, "Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Application", in *Proc. of HLT 2001*. San Diego.

[3] F. Metze, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schutz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana and E. Pianta, "The NESPOLE! Speech-to-Speech Translation System", in *Proc. HLT 2002*. San Diego.

[4] S.L. Oviatt, "Multimodal Interactive Maps: Designing for Human Performance", *Human-Computer Interaction*, 1997, pp. 93-129 (special issue on "Multimodal interfaces").

[5] S. L. Oviatt, A. De Angeli, and K. Kuhn, "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction", in *Proc. of CHI '97*. ACM Press, New York, 1997, pp. 415-422.

[6] S. L. Oviatt, "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture", in *Proc. of CHI '99*, ACM Press, New York, 1999, pp. 576-583.

[7] S. Burger, L. Besacier, P. Coletti, F. Metze and C. Morel, "The NESPOLE! VoIP Dialogue Database", in *Proc. of Eurospeech 2001*. Aalborg, Denmark.

[8] S. Burger, E. Costantini, and F. Pianesi, *NESPOLE! Deliverable D5 - Study on Multimodality, part 1*, 2002. In NESPOLE! Project website: http://nespole.itc.it

[9] E. Costantini, S. Burger, and F. Pianesi, NESPOLE! Multilingual and Multimodal Corpus, in Proceedings of LREC 2002, Las Palmas, Spain.