Combining Multi-Party Speech and Text Exchanges over the Internet

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Laboratory
Odense University, Science Park 10, 5230 Odense M, Denmark
nob@nis.sdu.dk, laila@nis.sdu.dk

Abstract

Bilateral or group text chatting over the Internet has become a favoured pastime for many people across the world. Yet it would seem that, in general, text chat is a severely impoverished mode of on-line communication compared to, e.g., fully situated human-human spoken conversation, video conferencing, or even speaking over the telephone. This paper explores what happens when on-line multi-speaker conversation over the Internet is added to text chat, creating what may become a widespread mode of communication in the near future. The system used is called the Magic Lounge. Magic Lounge offers a multimodal combination of text chat and spoken conversation for meetings and other encounters among ubiquitous users who may join the communication from workstations, PDAs and WAP phones. In addition, the system has a series of meeting history tools which provide various forms of structure to the spoken and text chat records of the meeting as it unfolds and after the meeting. The paper presents rather clear-cut results on the respective communicative roles of speech and text chat from a series of user tests with the system in which different groups of users performed scenarios designed to explore the combined use of text chat and speech. The results reported may generalise to a wide range of applications which combine text and spoken information representation.

1. Introduction

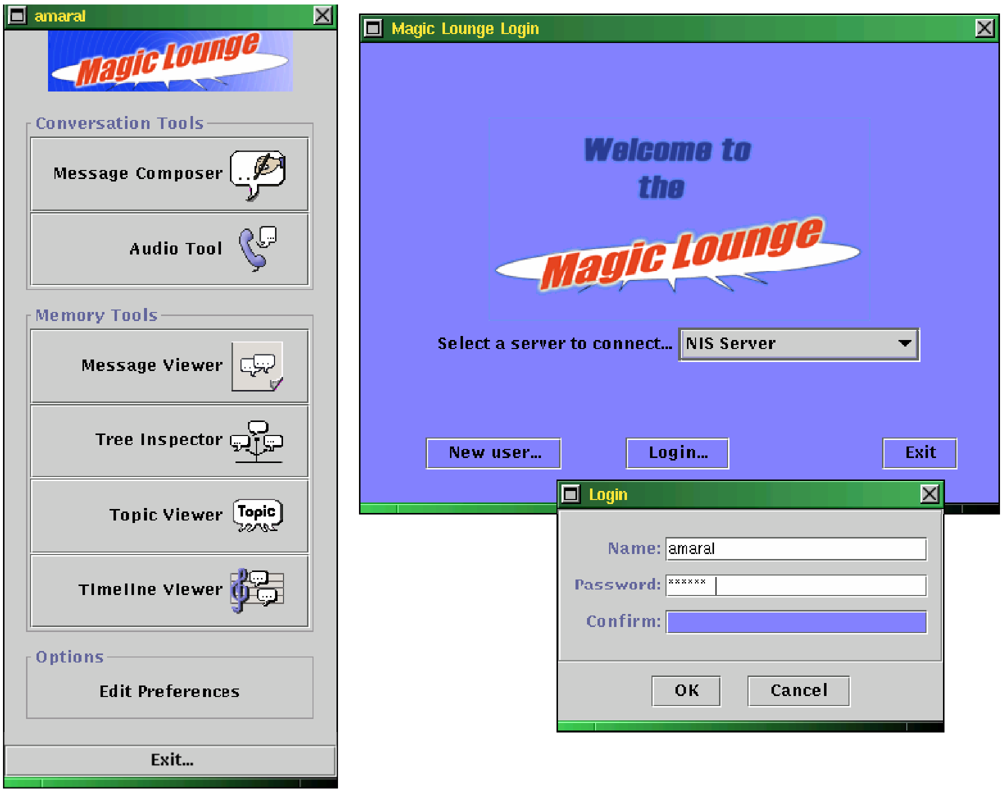
Text chat systems differ in their details but share the property that two or more users can exchange typed messages on-line and separately browse the chat record afterwards. We do not claim that text chat has not come to stay and that there are no purposes which are better served by text chat than by any other mode of human-human communication. For instance, the fact that chat is lacking in expressiveness, speed and informality compared to speech [1, 2] may be an advantage in some cases, such as when meeting someone for the first time on the Internet. However, soon we will all be able to simultaneously chat and speak together over the Internet, and this will certainly make up for the lack of expressiveness, informality, and speed which is characteristic of text chat-only. What will happen then? Will text chat more or less disappear because people will use speech instead when given the choice? Or will speech and text chat work so well together that, for most purposes, people will be using both? If they will, will they be using text chat and speech interchangeably or will these two very different modalities of linguistic communication tend to take on different and possibly complementary roles during communication? To seek answers to these questions, we did a series of user trials with the Magic Lounge system. The system is described in Section 2. Section 3 describes the user trials. Section 4 describes the meeting sub-tasks involved. Section 5 analyses the respective roles of text chat and speech. Section 6 concludes the paper.

2. The Magic Lounge

The Magic Lounge system has been developed in the i3 (Intelligent Information Interfaces) [http://www.i3net.org/] project Magic Lounge [http://www.dfki.de/imedia/mlounge/]. Magic Lounge consists of modules which enable multiple users to exchange labelled text messages, communicate via synchronous multi-party audio, and review previous and ongoing meetings through a structured memory. The functionality available varies according to the device on which the system is running. The desktop version of the system, running on a PC with a standard full-duplex audio card, enables all the above functionality. A PDA or a WAP phone enable only text messages and limited memory access. The memory and log facilities of the Magic Lounge are server-based. In order to join a session, users need to start the client and log on to the server. Users are asked to choose a password the first time they log in. Once a user is registered with a server, a toolbox pops up which offers the choice among a number of tools. Figure 1 shows the Magic Lounge Toolbox. The toolbox contains audio and text communication tools, memory access tools, and a preference setting tool.

The audio tool runs fairly independently from the text-based ones. Audio communication is supported by the real-time protocol (RTP) on top of IP multicast [3, 5], while the text part relies mainly on CORBA. At the moment, the connection between audio events and text messages in the memory is handled by a meeting browser or timeline viewer [4]. The timeline viewer provides the user with information on who is logged in (and for how long), and the distribution of communicative turns in terms of text messages and audio events sent per participant over time. With the audio tool, medium-sized user groups are able to communicate simultaneously in full-duplex mode while performing other activities in the Magic Lounge or using other tools on their desktops. Clicking on the ‘Audio Tool’ button automatically starts an audio session with a pre-defined multicast address.

The message composer enables users to write, reply to, topicalise, label, address, and send text messages to other Magic Lounge users logged on to the same server. The user may compose an entirely new message or reply to a previous message.



Select a server to connect... NIS Server

Name: amaral

Password: ******

Figure 1. The Magic Lounge Toolbox and Login Screen.

3. Methodology

Magic Lounge has been developed through a process of participatory design and iterative design and development [http://ravsted.nis.sdu.dk/video/]. The user group which has been collaborating with the developers throughout is composed of computer interested amateurs from smaller Danish isles. To achieve a broader representation of gender, computer literacy, and Magic Lounge (non-) proficiency, the user tests involved two other user groups as well. The first trial (two tests) involved Magic Lounge developers in order for them to evaluate the latest release of the system which had been developed in a distributed fashion at three partner sites. The second trial (one test) involved NISLab administrative staff who had little or no prior knowledge of the Magic Lounge. The third trial (two tests) involved Magic Lounge users from the Danish isles. The users in all three user groups knew each other well. In terms of user skills, the developers are skilled users of computer systems and of previous versions of the Magic Lounge. The NISLab administrative staff are standard office computer users and generally novice users of the Magic Lounge although one of the secretaries had tried a much earlier system version. The islanders are computer literates familiar with earlier versions of the Magic Lounge. As for user gender, the tests involved a total of three female and six male users.

The goal of the trials was to evaluate the Magic Lounge wrt. technical quality, observed functionality and usability problems, user satisfaction, and use of speech vs. text chat for different tasks. This paper focuses on the latter objective. All users used desktop machines or portables in an office environment. The individual test durations range from 30 minutes to 67 minutes. The logging of text contents was done directly in the Magic Lounge memory. The audio tracks of all meetings were recorded using an RTP (real time protocol) audio recorder [3]. The audio data gathered during the meetings was then converted to a standard audio format and written onto CD-ROMs in order to facilitate data analysis and make the data more widely available. Two of the three user trials were recorded on video.

The user tests were based on five different scenarios, each of which described a task which the users had to carry out together in the Magic Lounge. Thus, all user tests addressed task-oriented meetings and none of the tests addressed other kinds of encounters in the Magic Lounge, such as informal conversation. The tasks took into account the fact that the user groups had different levels of familiarity with the Magic Lounge. The developers trial included two scenarios, an on-line, free-form evaluation of the Magic Lounge and a website review task. The NISLab administrative staff trial included a single scenario specifying a party organising task. The islanders trial included two scenarios, a web browsing task to find a nice summer house for a German family of four and a questionnaire-based Magic Lounge evaluation task. Thus, two tasks done by two different user groups (islanders and developers, respectively) involved joint web browsing throughout, something which, on the one hand, is an obvious undertaking for people gathered in the Magic Lounge and which, on the other, poses potential “screen real estate” problems for users because the Magic Lounge itself includes a series of different windows for use during meetings. An introduction to the system was provided by a developer in the trials involving NISLab administrative staff and the islanders.

4. Tasks done using text chat and speech

We have already noted some of the variables in the user trials, i.e. computer proficiency, prior familiarity with the Magic Lounge, professional background, gender, and meeting task. A sixth variable is the meeting structure which varied from tightly structured, chaired meetings with an agreed agenda to ill-structured meetings. The Magic Lounge prototype worked sufficiently well throughout for the rich data from all tests to be taken into account in deriving the results below.

On the face of it, the participants carried out five different tasks during the trial sessions, i.e. the tasks specified in the test scenarios. In fact, however, those core meeting tasks should be considered sub-tasks addressed during the meetings in which the users carried out a number of other tasks as well. The full set of (sub-) tasks addressed may be listed as follows, indicating as well in which tests the sub-tasks were being addressed:

a) *exchange greetings* at the start and/or end of meetings (all tests);

b) test if the text and speech communication channels work (all tests);

c) *address technical, functionality and usability problems in actually using the software,* try to figure out how to operate the system, including calling an assistant or advising others to do so (all tests);

d) *address meeting organisation* more or less and with or without decision: present or create a meeting agenda, decide who chairs the meeting, who creates the meeting notes or minutes, or which core threads (topics) to use in text chat, propose to end the meeting, etc. (all tests);

e) *address (scenario-based) core meeting tasks:* informal system evaluation (Test 1), website design (Test 2), party planning (Test 3), holiday house offers (Test 4), and system evaluation questionnaire (Test 5), including in-point proposed ideas, counterproposals, arguments, motions, etc., tangential out-of-(core) tasks, and inconclusive discussions, checking whether the others have read the scenario, discussing how to interpret the scenario, exchanging URLs and other references, fragments of text, price information etc., guiding web navigation, discussing web sites, explaining abbreviations (all tests);

f) create meeting notes or minutes on the core meeting tasks (all tests);

g) *joke* about the core task or otherwise (all tests);

h) *describe what is going on in the communication right now,* including comments on the chat text as it is being produced (Tests 1,2,3,5);

i) review and comment on points in the text record at the end of the meeting (Tests 1,4);

j) summarise the discussion for a participant who has been absent (mentally or otherwise) (Test 3).

Interestingly, seven of the italicised phenomena (a through g) above were found in all tests no matter whether the users are novice users or not, whereas one was found in four tests (h), one in two tests (i), and one in a single test (j).

Most of the italicised phenomena above are found in face-to-face meetings as well, except for (b) and (c) which are crucial to virtual meetings. Even (c) has many counterparts in face-to-face meetings which make use of supporting technology. Other important points of difference to standard face-to-face meetings include the joint creation of meeting notes or minutes (f), spoken comments on the meeting notes or minutes as these are being produced (g), and the reviewing of the text meeting record at the end of the meeting (i).

Timewise, (a)-type phenomena occurred at the beginning and end of meetings, (b)-type phenomena in the beginning, (c) and (d)-type phenomena mostly in the beginning, (e) and (f)-type phenomena mostly after the initial phase and till the end of meetings, (g) and (h)-type phenomena at any time, and (i) and (j)-type phenomena towards the end of meetings.

In terms of user computer literacy, the only clear difference found between novice and skilled users is that the *frequency* of type-(c) phenomena (technical, functional and usability problems) was much larger in the test involving novice users (Test 3).

5. On the combined use of text chat and speech

In principle, all users might have decided to just make use of one modality, be it text chat or speech. In actual fact, they all used both modalities during their virtual meetings. Moreover, the data supports a series of generalisations on the respective use of speech and text chat in multi-party virtual meetings. In presenting those generalisations we refer to the list of meeting sub-tasks (a) through (j) in Section 4.

(a) + (b) *Exchanging greetings and testing if the text and speech communication channels work:* in all meetings, text chat/speech were used to make sure that everybody could send and receive text/speech messages.

(c) *Address technical, functionality and usability problems in actually using the software:* the test data shows that, rather obviously, users who are unfamiliar with the system or who are tasked to explore it spend some time sending test messages to explore the text chat functionality and making sure that they have understood the functionality properly. However, speech was used throughout to discuss technical, functionality and usability problems.

(d) *Address meeting organisation:* apart from the meeting agenda which, when present or created, was represented as text, virtually all meeting organisation was done through speech.

(e) *Address (scenario-based) core meeting tasks:* in the ill-structured meetings, especially in Test 3, some proposals and counter-proposals were made in text chat. Apart from that, virtually all proposed ideas, counter-proposals, arguments, motions, etc., tangential out-of-(core) task discussions and inconclusive discussions, checking whether the others have read the scenario, discussing how to interpret the scenario, guiding web navigation, discussing web sites, and explaining abbreviations were made through speech. Text chat, on the other hand, was used for exchanging URLs and other references, fragments of text, price information etc., i.e. information for which the exact wording mattered.

(f) *Create meeting notes or minutes on the core meeting tasks,* possibly including decision points. Text chat was used throughout.

(g) *Joke* about the core task or otherwise. With few exceptions, speech was used throughout.

(h) *Describe what is going on in the communication right now,* including comments on text as it is being produced. Speech was used throughout.

(i) *Review and comment on the points in the text record at the end of the meeting:* speech was used throughout.

(j) *Summarise the discussion for a participant who has been absent* (mentally or otherwise). Only speech was used.

With regard to the relative volume of speech and text chat in the trials, a total of 205 text messages (5780 words) was stored over 5 meetings compared to a total of about 4 hours of audio. We have not transcribed the spoken meeting contributions and hence cannot quantify the number of spoken messages exchanged. However, it seems clear that speech was used massively during the trials whereas text chat was being used much more judiciously. Only in the questionnaire-based test (Test 5) was the amount of text messages comparable to the amount of spoken utterances. In this test, the questionnaire acted as agenda, the meeting organisation worked perfectly, very few technical, functionality or usability problems occurred, and the islanders spent most of their time answering the 20 questionnaire questions in parallel.

The generalisations derived from the data analysis reported above are the following:

1. Virtual multi-party meetings in which participants communicate through speech and text chat are likely to include components (a) through (j) above. In particular, components (a) through (g) are likely to occur in all meetings.

2. Components (a) through (j) occur in partially ordered sequence (cf. Section 4).

3. Except for meeting tasks which by their nature demand that the participants all write throughout (cf. Test 5), speech is the preferred all-round communication modality.

4. When speech is available, text chat tends to assume particular, highly specialised roles. These are: (4.1) the obvious role of making sure that the text channel works and that the way it works is understood by the participant; (4.2) presenting the meeting agenda (if any); (4.3) exchanging information for which the exact wording matters; and (4.4) creating meeting notes or minutes on the core meeting tasks.

5. With minor exceptions, speech is being used for everything else. The roles of speech are: (5.1) the obvious role of making sure that the speech channel works and that the way it works is understood by the participant; (5.2) discussion exchanges of all kinds, on technical, functionality, and usability problems, meeting organisation, the core meeting task, tangential out-of-core task issues, interpretations, explanations, and text chat reviews; (5.3) joking and commenting on the current situation; and (5.4) summarising for absentees.

6. Conclusion

We have presented results from user tests of a multi-party virtual meeting system in which participants had the choice of communicating through text chat, speech, or both in order to solve scenario-based tasks. The results unambiguously demonstrate that the participants chose to use both text chat and speech to solve their tasks. Moreover, their use of speech and text chat show a clear pattern in the complementary roles assumed by the two modalities. Roughly but basically, text chat is used to structure the meeting through the agenda text, if any, exchange information for which the exact wording matters, and create a meeting record for posterity. Speech is used for discussion and related situated communication.

Given the fact that there is a human in the loop in both cases, we believe that it is possible to transfer the results reported from computer-mediated human-human communication to human-system communication. Increasingly, spoken language dialogue system developers need to consider how to integrate their speech-only input/output technology with other modalities for representing information. One such family of multimodal communication systems combines spoken input/output with static graphics (screen, display) output, such as text, images, graphs etc., for use in, e.g., mobile phones or cars. The results reported in this paper may be useful in addressing the puzzle concerning the possible roles of text output in multimodal systems belonging to the family of systems described.

Acknowledgements: The research reported in this paper was carried out under contract (1997-2000) with the European Commission's Unit for Future and Emerging Technologies and as part of the Proactive Initiative in Intelligent Information Interfaces (i3). We are grateful for the support. We would also like to thank our partners in the Magic Lounge project: LIMSI, Paris, France, DFKI, Saarbrücken, Germany, and Siemens AG, Germany.

7. References

[1] Bernsen, N.O., "Towards a tool for predicting speech functionality", *Speech Communication, 23:181-210*, 1997.

[2] Galegher, J. and Kraut, R., "Computer-mediated communication and collaborative writing: media influence and adaptation to communication", *Proceedings of the Conference on Computer-Supported Cooperative Work*, NY: ACM, 155-162, 1992.

[3] Luz, S. and Gromov, A., "Multicast audio conferencing and recording", *Technical report*, NISLab, Odense, Denmark, 1999.

[4] Roy, D. and Luz, S., "Audio meeting history tool: Interactive graphical user-support for virtual audio meetings", *Proceedings of the ESCA workshop: Accessing information in spoken audio*, 107-110. Cambridge University, 1999.

[5] Schulzrine, H., Casner, S., Frederick, R. and Jacobson, V., "RTP: A transport protocol for real-time applications", *IETF Internet Draft* (draft-ietf-avt-rtp-new-04), 1999.