IMMERSION IN SLIDEWORLD: A METHODOLOGICAL CASE STUDY FOR TECHNOLOGY SELECTION AND EVALUATION

Olivier Martinot, Arnaud Gonguet, Sylvain Squedin

Alcatel-Lucent Bell Labs France, Route de Villejust, 91620 Nozay - France
{Olivier.Martinot;Arnaud.Gonguet;Sylvain.Squedin}@alcatel-lucent.com

ABSTRACT

In the context of SlideWorld, a research project aiming at creating an immersive experience in videoconferencing, two video signal processing technologies have been developed and evaluated. A Smile Detector is used to increase the feeling of social presence and a Keyword Extractor allows focusing the attention on the video message. Those technologies have been evaluated for their intrinsic performances, but also in their contextual use in immersion with respect to users' feedbacks.

Index Terms— Keyword Extraction, Smile Detection, Video, Immersion

1. INTRODUCTION

SlideWorld is an Alcatel-Lucent Bell Labs' research project, aiming at creating an immersive experience for end-users during videoconferences. Our idea is to identify the key moments of the videoconference and to emphasise them in order to maximize the attendees' attention. To do so, some signal processing technologies had to be identified, contextualized and evaluated. In this paper, we present some sociology and cognitive psychology state of the art that enable us to define the high level objectives of those technologies (see §2). Then an early evaluation of the users' needs is presented, using a qualitative methodology from ergonomics (see §3). The selected signal processing technologies –*smile detection* and *keyword extraction*- are presented (see §4) and evaluated in order to refine the way they are used in *SlideWorld* to support the immersive experience (see §5). Some further enhancements for the technologies and their use in an immersive videoconference system are presented in the conclusion (see §6).

2. THEORITICAL BACKGROUND

Videoconferences systems are interactional artifacts [1] using gestures and facial expressions as communication strategies [2]. But in the Internet mediated communications one does not interact with another, but instead interacts with a representation of another, created by the technical system, what decreases the feeling of social presence [3][4]. The non-verbal word-gesture relation is also weakened due to the constraints of camera positioning [5]. To overcome those difficulties, one proposition is to use the theory of cognitive attractors to reinforce the feeling of presence by increasing the saliency of the information [3].

More specifically, immersion in a communication application, as defined in Bell Labs, is a mix of focused attention on the message and increased feeling of social presence. In cognitive sciences, the former is related with *cognitive immersion*. Both the *cognitive immersion* and the feeling of social presence are positively correlated with the user general satisfaction in a learning context [6][7][8]. The *cognitive immersion* relates to the expression *cognitive absorption* described "as a state of deep involvement or a holistic experience an individual has with an IT" [9], as well to the "Flow" experience characterized by a maximal mental state of immersion and focus [10]. The social presence has been defined as "being with others" [11], "level of awareness of the co-presence of another human, being or intelligence" [12] and "the degree of salience of the other person in the interaction" [13].

Based on this review of the sociology and cognitive psychology state of the art related to immersive communications and videoconference systems, our goal for *SlideWorld* has been to identify technologies for detecting the salient non-verbal information about attendees to increase the feeling of social presence and to find a way to maximize the attendees' attention.

3. USER NEEDS ANALYSIS

We had to ground our software engineering research on empirical evidences, because an immersive videoconference system such as *SlideWorld* cannot be a techno-push innovation, and require a co-construction with users [14]. When there is little research conducted on a particular phenomenon, or where research hypotheses require increased focus, an evolutionary perspective is recommended: An initial exploratory study gathering qualitative data is undertaken, to explore a wide range of

topics. The collected data is then analyzed, and the important findings from this initial study are refined and used as hypothesis for further studies [15].

We organized a qualitative evaluation of the *SlideWorld* concepts, with 9 participants familiar with videoconference systems. The first step was to present the issues that we think should be addressed in *SlideWorld* (increase the level of social presence and the users' attention to the message) and to collect feedbacks. The validity criterion in such a methodology is the answers' convergence.

In order to reduce the social isolation of the remote audience, the test participants expect that the application will at least give them the list of the attendees to the meeting (P2, P6), or in the best case some means to interact with the other attendees or with the presenter (P1, P5, P8, P9). However, this social *isolation* is considered as an advantage by the participants (P2, P3, P4, and P7) and some of them are seeking for it (P3). Indeed, the social pressure being reduced (P2, P3, P4, P7, P8), it is then possible to have other activities during the presentation (P2, P3) or even to leave the presentation when it is not interesting any more (P4, P8).

As a conclusion, it seems that to increase the feeling of social presence, a non-intrusive solution should be investigated, *preventing* from direct interactions with other attendees while increasing the "level of awareness of the co-presence of another human".

Few expectations were provided by the test participants about the boredom of a presentation. Only 3 of them expressed a need to interact with colleagues or with the presenter in this specific case (P6, P8, and P9). Moreover, the problem is a proper problem for 2 of the test participants only (P1, P2), that explain that during a videoconference there is nobody in front of us to talk directly to us, to stare at us, to captivate us. For others, the boredom problem exclusively comes from the presentation and thus the application could hardly overcome the problem (P3, P4, P7, and P8). This problem is also limited for videoconferences because before the presentation starts the attendees perform other activities (P1, P2) or get connected late (P1, P4). Moreover, during the presentation, the remote attendees can do other things than listening, or even disconnect from the application if the presentation is getting too boring (P2, P3, and P8).

Increasing the users' attention on the message should then be achieved with either optional means (when users think that the presentation is unclear) or by increasing the focus on the message thanks to entertaining video edition (when users don't want to focus on the message).

4. TECHNOLOGIES

4.1. Smile Detector

The *Smile Detector* (Figure 1) informs the auditors about the other auditors' interest by detecting their face expression. This technical component exploits the Active Appearance Model (AAM) proposed by Cootes et al. in 1998 [16]. Among different applications, AAM is used to model face expression through a statistical model of shape and appearance. A training phase on a set of images allows creating a set of model parameters from landmark points. The added value of AAM lies in the compactness of the model parameters. Then during the applied phase, specific parameters values related to the smile are detected. Based on the state-of-the-art, this smile detection is performed in real-time [17].

Figure 1: Smile Detector control pane

In *SlideWorld* the outputs of the *Smile Detector* are used to animate abstract avatars of the attendees in order to increase the feeling of social presence in an anonym way, in order to by as less intrusive as possible.

4.2. Keyword Extractor

For the auditors, a *Keyword Extractor* was developed to identify the meaningful keywords extracted from the previous minute's speech of the videoconference discussion. Then those keywords are displayed in the video or in an on-demand tag cloud. The objective is to sustain the auditors' attention in the case where attention is disturbed by a call or an email, or if the message is unclear, by allowing the auditor to reconnect to the current speech. Several steps are necessary to reach that objective (Figure 2). Firstly, based on silence detection or on a timeout, the audio stream is cut into sub parts that allow a continuous analysis of the speech. Then speech-to-text step is performed using on-the-shelf ASR (Automatic Speech Recognition) tools, Nuance [18] or Jibbigo [19] ones, in order to get a time-stamped text transcript of the speech. Then a keyword extraction is performed that uses text manipulation techniques like POS tagging and TF-IDF (term frequency–inverse document frequency) method. In a last step, in order to reduce errors due to the speech-to-text engine or keywords extraction, those keywords can be semantically processed for consistency consolidation in real-time, relying on internal or contextual information.

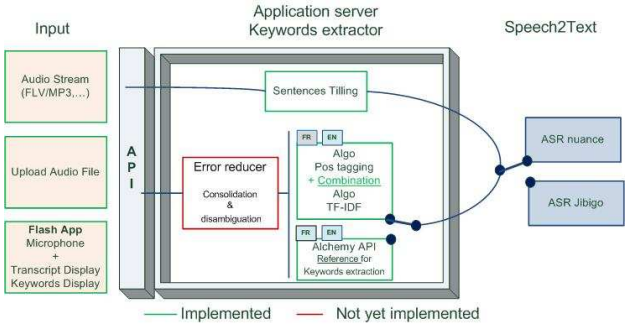

Figure 2: Architecture of Keyword Extractor

The part-of-speech tagging (POS tagging or POST) is a grammatical tagging algorithm which has the ability to assigns properties, or parts of speech to each word (and other tokens), such as noun, verb, adjective, etc [20]. The expected result provided by the POST is the possibility to choose specific tags like only the noun, to extract proper noun or create new properties with couple of tags: subject, associated verb and direct object of the verb. In our case, the keywords extraction is performed using either Alchemy API [21] or using combination of POS tagging and TF-IDF model [22].

4.2.1. Evaluation data and methods

The *Reference Discussion Test* in our study is an audio stream corresponding to an open discussion on Google+ with 4 peoples during about 62 minutes. Each people have different audio quality, and several audio artifacts appear during the discussion, like simultaneous speaking or audio feedback (also known as the Larsen effect). A manual transcript was performed to build the ground truth. In this transcript, 7947 words were identified.

The first evaluation aims to compare the results of the speech-to-text step for several tools. The manual transcript of the *Reference* Discussion *Test* is analyzed and compared with the automatic transcript done by the ASR of Jibbigo (ASR1 case) and Nuance (ASR2 case). The objective is to select only ASR tools that reach minimum performances. In the second evaluation, the list of keywords is extracted for each transcript with one of the POS tagging algorithms: Alchemy API (POST1 case) and our own implementation of POS tagging (POST2 case). Then, the cosine similarity computation is used to compare the both extracted keywords lists. It measures the similarity between two vectors by measuring the cosine of the angle between them.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

For both evaluation steps, if the results are reaching a predefined level, we will then be able to apply several texts mining algorithms like text categorization, text clustering, and concept/entity extraction. A threshold value was chosen for the cosine: 0.8. It gives for instance, (1,2,**3**,4,5) vs. (1,2,**6**,4,5) = 0.8.

4.2.2. Evaluation results

For the first evaluation, Table 1 shows the results of each of the tested ASR tools, compared with the manual transcript: ASR2 nearly reaches the threshold. This is probably caused by a higher Jibbigo sensitivity to the input audio quality and to the sound artifacts. We will thus use the Nuance ASR tool in our *Keyword Extractor*.

ASR	Cosine Similarity
ASR1	0.37772107786846
ASR2	0.7991760797989

Table 1: Reference Discussion Test ASR tools results

In Table 2 is shown the ASR2 speech transcript between 28mn and 30mn 15s. The second evaluation is performed only on the ASR2 results. In Table 3 are the results of the each of the keyword extraction algorithms for the speech between 28mn and 30mn 15s.

[00.28.00] She use a common tag across both streams simultaneously sort out to instantiate it and then later you can blend them together because they both absorbers synchronized you know metadata hot |

Rectitude would share so to me I guess the general followed be harder you get our shared metadata object attribute does one of these tags being aligned with different lists streams you have indirectly you of the |

[00.28.30] Like is a time is a reasonable one maybe that's all we need for a lot of things so some of the integration would would be fancy because when one thing we were doing even like right now for example because we can't read the power

Point were clicking each of us individually is clicking through PowerPoint and if over a time stamping Riyadh we had a Journal that showed what are no image was during you general then we could play it |

[00.29.00] Back and say well this is what I was seeing when we are having this part of the conversation etc guard this is our made remarks one and I was visiting this webpage and read reading this in a while W3C site or something | Is that I you're all that is fine

That getting it back yet what you're saying about I'm I three it is right now I have one I put times and I was thinking that it would be of use |

[00.29.45] One I'm using now Zarrella but I'm beginning of our so that you position basically within that narrow a the back of the presentation but in order to say

But you're saying I was thinking about having by absolute on and out of basically say no coordinated UTC whatever life the and now it artifacts are being produced in **[00.30.15]**

Table 2: ASR2 transcript for the selected period

Start date	End date	Extracted keywords POST1	Extracted keywords POST2
00.28.00	00.28.30	Common tag ; metadata ; different lists ; Rectitude ; blend ; tags	tag; Rectitude; metadata; attribute
00.28.30	00.29.00	reasonable ; Riyadh ; PowerPoint ; general ;	Riyadh; Point; PowerPoint; Journal
00.29.00	00.29.45	W3C site ; webpage ;	Conversation ; etc ; guard; webpage ; W3C
00.29.45	00.30.15	Presentation;	Zarrella ; UTC

Table 3: Keywords extracted in the selected period

To conclude based on those results, POST2 algorithm gives meaningful results that are comparable to the POST1 algorithm. The envisaged next steps are to improve the POST2 algorithm with the implementation of the consolidation process and with up to date TF-IDF corpus, clustered by topics.

5. USER EVALUATIONS

Following the qualitative methodology presented in §3, we had a second step of SlideWorld evaluation with 9 participants. We presented the technical solutions to the identified problems (increase the feeling of social presence with the Smile Detector and increase the attention to the message using the Keyword Extractor) and asked the participants to provide feedbacks and to propose enhancements.

The *Smile Detector* use to animate abstract avatars of the remote attendees allows reducing the feeling of isolation (P1, P2, P3, P5, P6, and P7). Moreover, being seen by other attendees strengthens the links between attendees (P4, P6, and P8). However, because of the social pressure, 2 participants (P5, P9) – when they were explained how the system should work – found themselves very uncomfortable with the system implicitly capturing their mood and presence. Finally, 3 users noted that the audience representation using abstract avatars would not allow representing a huge number of remote attendees (P2, P7, and P9).

It was suggested that allowing attendees to send temporary signals such as "I don't understand" (P1) or "it's boring" (P5) would allow providing feedback from the audience while respecting those who don't want to be spied (P5, P9).

The tag cloud based on the *Keyword Extractor* allows joining attendees (or distracted ones) to instantly catch up with the context of the presentation (P3, P5, P8, and P9). On the other hand, the tag cloud is seen as useless by the attendees that follow the presentation (P1, P2, P3, P4, P7, P9), because the provided information is not enough precise (P2) and is redundant with the presenter's speech (P1, P3, P4).

It was suggested to be able to scroll through the tag cloud by 5 minutes periods (P9) in order to catch up more easily with the presentation.

6. CONCLUSION

This paper shows how a set of signal processing technologies have been selected and evaluated in the specific context of immersive videoconferencing. Our idea was to identify the key elements of the videoconference in order to emphasize them. To achieve immersion, the sociology and cognitive psychology state of the art teach us that the feeling of social presence and the focused attention on the message are key elements. An early evaluation of the users' expectations showed that the situation of videoconferencing is related with a sense of privacy that guided us toward non-intrusive technical solutions. Moreover, because we somehow wanted to maximize the attendees' attention on the message against their own will, this had to be handled smoothly, by enhancing the overall quality of the presentation.

The *Smile Detector* detects the face expressions and can be used to animate abstract avatars of the remote audience with some mood information. The *Keyword Extractor* has been developed and evaluated to identify meaningful keywords in the presenter's speech in order to enhance the video message or to allow distracted attendees to catch up with the presentation. Nuance has been selected for its transcription quality and its performance on our reference audio file. It's difficult to determine the failure origin during the transcription test because the ASR is a black box, but in our case and according to the tests done, Nuance seems to have the best sound artifacts tolerance and allows to propose our own keywords extractor solution which has proven to be efficient enough compared to the Alchemy API. Next step will be to improve our algorithm with the implementation of the consolidation process and with up to date TF-IDF corpus, clustered by topics.

Finally, the user evaluation of the proposed solutions allowed us to confirm the interest of the *Keyword Extractor* for inattentive attendees, though we know that meaningful

keywords should be smoothly integrated in the video message, to create a more entertaining video, in order to directly sustain the users' attention. The use of the *Smile Detector* will probably raise privacy concerns, and anonymizing the results seems to be critical. The next step of this study will be to rework the integration of the developed technologies in order to better fit the users' expectations and the design objectives of increasing the feeling of social presence and attention to the message.

7. REFERENCES

[1] M. Fornel, "Le cadre interactionnel de l'échange visiophonique", In Réseaux, volume 12, 64, 1994, pp.107-132.

[2] B. Bonu, "Connexion continue et interactions ouvertes en réunion visiophonique", In De la rue au tribunal : Etudes sur la visiocommunication, C. Licoppe, Réseaux, Vol 25 2007 /144.

[3] S. Lahlou, "L'activité de réunion à distance," In De la rue au tribunal : Etudes sur la Visiocommunication, Réseaux, Vol 25, 2007/144.

[4] N. Curien and M. Gensollen, "La communication dans un groupe de travail : analyse du fonctionnement interactif et le marché des téléconférences", Réseaux n°10, février 85, Téléconférence, 1991.

[5] C. Licoppe, "De la rue au tribunal : Etudes sur la visiocommunication", In Réseaux, vol 25 2007 /144.

[6] C.N. Gunawardena and F.J. Zittle, "Social presence as a predictor of satisfaction within a computer-mediated conferencing environment", The American Journal of Distance Education, 11(3), 1997, pp. 8–26.

[7] K. Swan and L.F. Shih, "On the nature and development of social presence in online course", Journal of Asynchronous Networks, 9, 3, 2005, pp. 115–136.

[8] R. Saadé and B. Bahli, "The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: an extension of the technology acceptance model", Information & Management, 42, 2, 2005, pp. 317-327.

[9] R. Agarwal and E. Karahanna, "Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage", MIS Quarterly, 24, 2000, pp. 665-694.

[10] M. Csikszentmihalyi, "Flow: The Psychology of Optimal Experience", Harper Perennial, New York, 1990.

[11] C. Heeter, "Being There: The subjective experience of presence", Presence, 1(2), 1992, pp. 262-271.

[12] F. Biocca and K. Nowak, "Plugging your body into the telecommunication system: Mediated embodiment, media interfaces, and social virtual environments", In C. Lin and D. Atkin (Eds.), Communication technology and society, Hampton Press, 2001, pp. 407-447.

[13] J. Short, E. Williams, and B. Christie, The social psychology of telecommunications, John Wiley & Sons Ltd, London, 1976.

[14] B. Latour, "L'impossible métier de l'innovation technique", Encyclopédie de l'innovation, in Mustar P., Penan H. (eds.), Paris, Economica (2003), pp.9-26.

[15] M. Wood, J. Daly, J. Miller, and M. Roper, "Multi-Method Research: An Empirical Investigation of Object-Oriented Technology", Journal of systems and Software, vol 48, 1, 1999, pp. 13-26.

[16] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models", European Conference on Computer Vision (1998).

[17] D. Govindaraj, Application of Active Appearance Model to Automatic Face Replacement, Eurographics, EG2011.

[18] http://www.nuance.com

[19] http://www.jibbigo.com

[20] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A Practical Part-Of-Speech Tagger", In proceedings of the Third Conference On Applied Natural Language Processing, 1992.

[21] http://www.alchemyapi.com

[22] M. Ribière, J. Picault, and S. Squedin, "The sBook: towards social and personalized learning experiences", Proceedings of the third workshop on Research advances in large digital book repositories and complementary media, 2010.