AUGMENTING ACCESSIBILITY IN SOCIAL NETWORKS
A Virtual Presenter

Leonelo D. A. Almeida[1], Elaine C. S. Hayashi[1], Júlio C. Reis[1,3], Paula D. P. Costa,[2]
M. Cecília C. Baranauskas[1]and José Mario De Martino[2]
[1] *Department of Information Systems, Institute of Computing, University of Campinas*
Av. Albert Einstein, 1251, 13083-970, Campinas, SP, Brazil
[2] *Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering*
University of Campinas, Av. Albert Einstein, 400, 13083-852, Campinas, SP, Brazil
[3] *Center for Information Technology Renato Archer. Rodovia Dom Pedro I, km 143,6, 13069-901, Campinas, SP, Brazil*

Keywords: Inclusive social networks, Talking heads, Facial animation.

Abstract: *Vila na Rede* is an Inclusive Social Network (ISN) system developed as a product of *e-Cidadania's* Project, with the objective of being accessible to a broad variety of users, including those less familiar with technology or with low literacy skills. Different features were incorporated into *Vila na Rede* in order to provide its users with scaffolding learning support that helps them profit more from the system. One of these features is the Virtual Presenter, a speech synchronized facial animation system integrated to a text-to-speech synthesizer (TTS) that allows users to have textual information alternatively presented by a talking head. This paper investigates the integration of the Virtual Presenter into *Vila na Rede* system, and discusses results of activities conducted with the target audience to evaluate this new feature at the ISN.

1 INTRODUCTION

E-Cidadania (2011) is a Brazilian research project that has taken the challenge of developing systems that allow the access to the Information Society, contributing to the constitution of a digital culture and respecting the diversity of the population. As a result, the project has launched *Vila na Rede* (2011), an Inclusive Social Network (ISN) system, designed with, by and for Brazilian people. *Vila na Rede* is being used by citizens all over Brazil and has even been accessed from abroad, announcing ideas, goods and other initiatives from different communities.

The design of systems that make sense and that are accessible to citizens requires a rather socio-technical view of the problem. Because of that, the research has adopted as a methodological reference, the principles and concepts of both Participatory Design (Mumford, 1964; Schuler and Namioka, 1993) and Organizational Semiotics (Stamper *et al.*, 1988). *E-Cidadania's* researchers have been successfully involving end users and developers in Semio-Participatory Workshops (Neris *et al.*, 2009, Hayashi and Baranauskas, 2009) in order to design and construct the ISN.

In developing countries such as Brazil, India and China the access to knowledge is still restricted to a small portion of the population. Creating Web applications that can aid people better and easily understand and access information is necessary and a challenge to be faced. To mention a few statistics, as stated by the Brazilian Internet Steering Committee (2009), 47% of the population has never used the computer and 55% have never accessed the Internet. It points out that most of the considered target population for ISN is inexperienced with computers.

Moreover, regarding literacy skills, the Indicator of Functional Literacy in Brazil (Institute Paulo Montenegro, 2009) (INAF in its Portuguese acronym) points out that in 2009, 27% of the Brazilian population between 15 and 64 years old were considered functionally illiterate, defined as the population with less than 4 years of schooling that are not able to perform simple tasks involving words and phrases. Using a broader concept of functional illiteracy, according to the same Indicator, the majority (52%) of the Brazilians in the same described situation reaches no more than the degree of rudimentary literacy, *i.e.*, they only have the

D. A. Almeida L., C. S. Hayashi E., C. Reis J., D. P. Costa P., Cecilia C. Baranauskas M. and De Martino J..
AUGMENTING ACCESSIBILITY IN SOCIAL NETWORKS - A Virtual Presenter.
DOI: 10.5220/0003471502250234
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 225-234
ISBN: 978-989-8425-56-0

ability to locate explicit information in short texts or make simple math. Thus, they are not able to understand longer texts, and even worst, 9% of these individuals can be considered absolutely illiterates. These statistics can illustrate the social reality of the target community aimed to be included. Nevertheless, designing systems for the inclusion of these people requires multidisciplinary expertise, and adequate technology solutions that can provide barriers-free access to information to every citizen.

In workshops involving our target users, we have understood how the figure of an avatar in the system would be beneficial (Hayashi and Baranauskas, 2008) to support the less experienced users. Those workshops had the objective of analyzing how users make sense of different multimedia outputs to retrieve information. Users were grouped to perform a task in a role play situation in which they had to find the answer for a question. Each group was exposed to a different way of finding the answer. One group had access to the information in written format. Another group was able to listen to the information, as if they were facing an Automated Response Unit (ARU), like those found in call centers. The third group had the same information available in images. The last group consulted the information from a real person, like in an information desk. The group with best results was the last group. During the discussion with all groups, they reported their wish to have similar attendants in the systems – *i.e.*, the figure of a person to support them in their online tasks.

In another activity, we found that computer synthesized voices would probably be well accepted by our target users. In that experiment, which was in fact related to another project and scope, users thought that the synthesized voice was the recording of the voice of a foreign researcher (speaking in Portuguese the users' mother language, but with accent). They enjoyed the voice and had fun with it, and most important, they were able to understand it. During yet another encounter, users accidentally had access to a feature that was being built at *Vila na Rede's* test environment, and that was not ready for use. This feature – as it was implemented at that time – was able to synthesize text extracts into speech. Even though the final objective of this feature was different, since users were excited about the possibility of having the text read for them, we decided to implement the feature as it was. Such new feature can be referred in the literature as a "talking head" (TH).

For an ISN context, considering the target users mentioned, a TH that could read the digital contents (*e.g.* long announcements) and that could be easily available for people use would be appreciated. The main advantages of a TH are that it facilitates the access to information, is ready to use in the same environment where the information is found, and it delivers the information via synthesized speech in a more friendly way, through the avatar. Even knowing that an implementation of a TH in an ISN may not look exactly as a real person, we consider such approach an import attempt to augmenting accessibility in ISN. TH seems relevant for an ISN since it can have an important role in helping functional illiterate people to understand contents. Such investigation is also relevant since there is a restricted and limited TH offering for Portuguese language. This language brings a complexity to synthesized speech as to have it as close as possible to the human voice.

This paper is organized as follows: Section 2 presents a brief review on TH systems; Section 3 describes the process of implementing the existing system at *Vila na Rede*; Section 4 presents the Semio-participatory workshop in which the feature was tested and discuss the results from this experiment. Finally, Section 5 concludes the paper.

2 TALKING HEAD

Facial animation systems synchronized with speech, or THs, constitute an enabling technology for the development of human-computer interfaces capable of reproducing the natural and intuitive face-to-face communication mechanisms that we are well familiar with. These systems can be designed as an alternative to WIMP (Windows, Icons, Menus and Pointing devices) interfaces, simplifying human-computer interaction and increasing the accessibility to users without specialized training.

Convincing THs are able to turn applications more human, involving and attractive. Synthetic human faces can be used, for example, to implement virtual characters in the role of lecturers, tutors, assistants, newscasters, avatars, salespeople and customer care or user support attendants (Cosatto and Östermann, 2003; Johnson *et al.*, 2000; Östermann and Millen, 2000; Pandzic, 2002). In some studies virtual characters act in activities with children in emergent literacy and as a technological intervention tool for children with autism (Ryokai *et al.*, 2003; Tartaro and Cassell, 2006). Additionally, flexible face models can be adapted to build avatars for an individual. Finally, although broadband data access is being popularized, computer facial

animation is a model-based video-coding approach and can be an alternative to video transmission in severe situations of limited data traffic capacity.

Typically, there exist two main approaches for face modelling on TH systems: the image based, or 2D, approach and the geometric face modelling, or 3D, approach. Image-based, or 2D, facial animation manipulates photographic pictures from an image database that are captured from a real face, synthesizing the final animation through their appropriate processing, sequencing, concatenation and presentation. This approach inherently generates photorealistic animations due to photographic nature of manipulated images. On the other hand, advanced and modern 3D face modelling techniques are successful in synthesizing natural looking faces and high quality images of rigid movements of the head. However, the use of polygonal meshes or other geometric models to reproduce plastic deformations, like the mouth dynamics during speech or the details of human face visual appearance, requires sophisticated models and animation control strategies, implemented at high computational costs.

The TH system adopted at *Vila na Rede* is AnimaFace2D, a 2D facial animation system capable of being integrated to a text-to-speech synthesizer (TTS) for Brazilian Portuguese (Costa and De Martino, 2010). Figure 1 summarizes the expected inputs and outputs of AnimaFace2D synthesis process. The following characteristics make this system well suited for integration with *Vila na Rede*:

- AnimaFace2D is a 2D facial animation system designed for Brazilian Portuguese language, the same language of *Vila na Rede* users;
- The system provides realistic modelling of speech articulatory movements (Costa and De Martino, 2010);
- AnimaFace2D visually animates any speech content from a reduced image database of just 34 photographs visemes (the visual counterpart of a language phoneme);
- The system can be adapted to any available TTS system that provides timed phonetic transcription of synthesized speech;

The underlying principles of the facial animation synthesis methodology proposed in (Costa and De Martino, 2010) can be adapted to other languages and to different face models.

The integration of this system to *Vila na Rede* constituted the Virtual Presenter Core (VPC) of the proposed architecture of our ISN.

3 THE VIRTUAL PRESENTER AT *VILA NA REDE*

The architecture built to support the integration of VPC into our ISN is based on making virtual presentation videos as services for the ISN *Vila na Rede*. Figure 2 illustrates the architecture using two servers, the first hosting the ISN and the other performing the processing of VPC and hosting the resulting videos. When new or updated data is posted from the user interface to the ISN the text content is submitted to a text-to-speech server in this case Festival (University of Edinburgh, 2010).
The text-to-speech server provides the synthesized speech and its corresponding timed phonetic transcription file. After the synthesis process, the ISN makes a request to the Virtual Presenter Service (VPS).

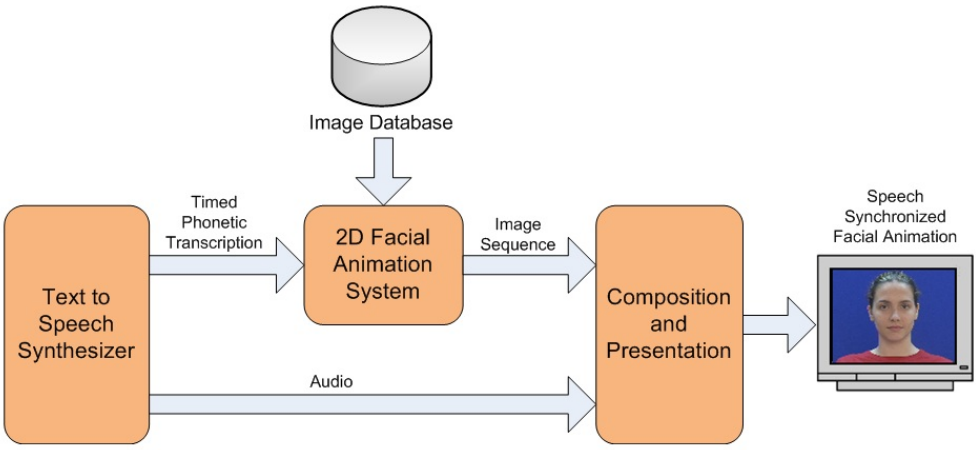

Figure 1: Synthesis process.

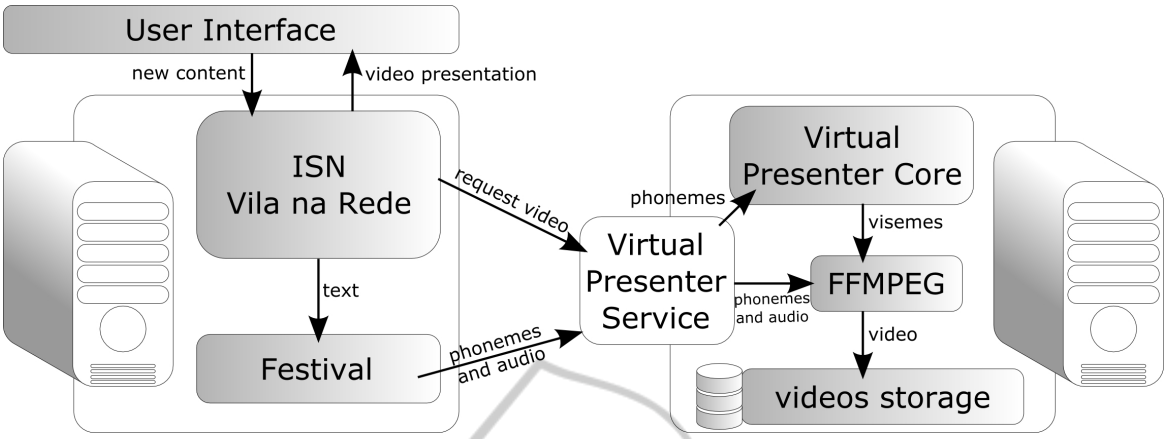

Figure 2: Virtual Presenter architecture integrated to the *Vila na Rede* ISN.

Figure 3: Virtual Presenter in the ISN Vila na Rede.

So VPS starts by getting the phoneme and audio files from the ISN server and calling the VPC. VPC is a C++ application that uses OpenCV (Open Computer Vision Library (2010). It uses a database of 34 images and a timed phonetic transcription file to synthesize a facial animation. Currently the VPC generate animations considering approximately 30 frames per second. Each frame is an image of 320 x 240 pixels. At end of the VPC execution, VPS calls FFmpeg (2010) audio and video library to mix facial animation and speech audio files, resulting in the final talking head video. Finally, VPS informs the ISN that the virtual presentation is available for the posted content.

In the user interface layer the process for generating video presentations is asynchronous to the users' interaction, so that users do not have to wait the end of the generation process to continue interacting with the ISN. When a person posts some content it is not necessary to explicitly request the video creation or wait until it is generated. The content types processed by the Virtual Presenter involve announcements of ideas, products, services and events. After the system finishes the video generation, an icon indicating the video availability is displayed along with the content, as presented in Figure 3A. By clicking on the icon the ISN loads the video and presents it at the right region of the website (see Figure 3B). To present the video the ISN uses the JWPlayer (Long Tail Ad Solutions, 2010).

By the time this research was developed there was not a free Festival extension for Brazilian Portuguese language that offered a female voice. To overcome this issue, we changed the pitch of the male voice to obtain a neutral tone. This was necessary due to the fact that the AnimaFace2D visemes database is composed of photos of a woman's face. This modification did not affect the length of the video nor the phonemes pronunciation

but made the voice sound more artificial than the original configuration.

4 A SEMIO-PARTICIPATORY EVALUATION OF THE VIRTUAL PRESENTER IN *VILA NA REDE*

In this section we describe the activities that were carried out in order to evaluate the new functionality regarding the virtual presenter at *Vila na Rede*. The three main aspects to be evaluated were: 1) The use of Festival, considering the quality of the voice, its timbre and speed of the speech; 2) The Virtual Presenter and the video image x speech audio synchronism, how close to reality the videos are, and the role of the video in helping to understand the audio; 3) How both (image and audio) were implemented at *Vila na Rede*, considering the size of the video, speed and the work done to incorporate it into the website.

The activity was conducted in a laboratory facility, at the University. Eleven members of the target community were invited to participate in the experiment. They all have residence in Campinas city, but they came from different regions of Brazil (namely north, northeast, south and southeast). Their ages ranged from 22 to over 60 years old. Considering the schooling level, 40% of the participants have university degree, 30% have high school degree while 10% have incomplete high school; 10% have just the elementary school education and 10% have not completed even the elementary school level. The experiment included people of different social profiles such as: housewife, cook, handicraftsman, hairdresser, seamstress, retired people, teacher, student and others. Concerning their experience with ICT or other electronic devices, all of them have at least one TV set at home; two have never used ATMs and other three do not like to use them. Only one person does not have a cell phone, all the others use the cell phone on a daily basis. Six of them use Short Message Service (SMS). None of them have Internet access in their mobiles. Two out of the eleven participants do not have a computer at home. Two of those who do have a computer at home do not use it. Those who do have a computer at home also have Internet access, being three of them dial up access. Only two participants were familiar with online social networks and part of at least one (Orkut).

In the activity for the evaluation of the virtual presenter, each participant was assigned either to one desktop computer or to one laptop; they all received head phones to listen to the audio and they were all able to see the videos. Each two participants were assigned to one researcher who had the role of observing their behaviour. The observers were in charge of playing the announcements in the order that was specified for the task.

Table 1: Announcements used in the activity and its form and content.

Id	Media	Content (in Portuguese)
Shorter announcements:		
SA	audio	"Ribbon embroidery, cross stitch and more. Negotiable prices" (in Brazilian Portuguese: "bordados fita, ponto cruz e outros. Preços combinar").
SB	audio + video	"'Fuxico' doll. This is a doll made of 'fuxico'" (in Brazilian Portuguese: "boneca de fuxico. Essa é uma boneca feita de fuxico"). Fuxico is a craft technique that uses pieces of fabric to decorate clothing, pillows and other utensils.
Longer announcements:		
LA	audio	"New functionality in the Vila: The Virtual Presenter. Soon we will have the Virtual Presenter that will read the announcements for you. The launch will be in March" (in Brazilian Portuguese: "Nova funcionalidade no Vila: a Apresentadora Virtual. Em breve teremos a Apresentadora Virtual que lerá os anúncios para você. O lançamento será em março").
LB	audio + video	"June Festival at SAMUCA. SAMUCA - 301 Antonio Provatti street - Jd. Triunfo. Come to participate of the SAMUCA's great festival, we will have typical booths and many gifts for everyone. Do not loose it" (in Brazilian Portuguese: "Festa Junina SAMUCA. SAMUCA Rua Antonio Provatti 301 Jd .Triunfo. Venham Prestigiar e Participar da Grande Festança do Arraia do SAMUCA, teremos barracas típicas e muito brindes para todos, não percam").

The announcements that were shown to participants were previously chosen by the researchers. The announcements were chosen according to the following criteria: number of words and phrases similar between the two announcements

of each type (*i.e.* short and long), avoiding texts that contain abbreviations or orthographic problems, and texts produced in the participant's context.

In total, participants saw and/or listened to four *Vila na Rede* announcements. Two were shown with video and sound and other two announcements had only the audio. This control was made in order to provide us with means to compare the level of comprehension to deduce the influence of the media.

From the four announcements presented to the users, two were rather short ones, intended to be reproduced word by word by the participants. The other two were longer announcements and participants were supposed to simply inform the overall idea, and not the literal transcription. Table 1 shows the announcements that were chosen and shown to the participants, as well as announcements with only the audio and with video and audio. Each announcement has one Id as "SA" that stands for "short announcement A" or "LB" for "long announcement B". The announcements' contents were originally written in Brazilian Portuguese (orthographic errors were omitted in the English version in Table 1).

Right after the task regarding the announcement, participants completed a form with the information asked. As mentioned before, for the first two announcements, participants were asked to write in this form the actual content heard; and for the last two, the general idea understood.

5 RESULTS

The results obtained from the forms were analyzed considering two metrics. First, considering the two shorter announcements, we counted the substantives and verbs from them *i.e.*, in the SA announcement there are 7 words and, for the SB announcement there are 5.

Second, as the participants were supposed to write what they had understood about the longer announcements we identified the topics in each of them. For the LA announcement there are 3 topics *i.e.*, it is a new functionality, it reads announcements, and it will be launched in March. For the LB announcement there are 4 topics *i.e.*, it is a festival, it will happen at SAMUCA, it is an invitation, and there will be food and gifts. Figure 4 presents the correct answers average by media

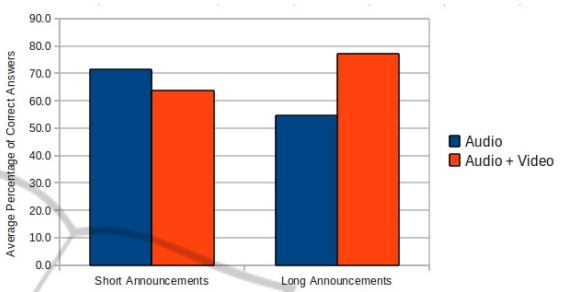

Figure 4: Correct answers average by media.

Considering that the objective for the longer announcements was to check the correct understanding of the media presentation instead of the exact words, we considered the number of topics mentioned in the responses provided by the participants. We only listed the correct data; participants were not supposed to write exact words but their understanding.

In order to evaluate the results, we considered the relative coverage of the correct responses for each announcement, using the set of individual responses, the mean, mode and standard deviation (see Table 2). Based on the mean we verified that, for the shorter announcements, participants were more precise using the audio version of the announcement. For the longer announcements the opposite occurred.

Further analysis and conclusions can be driven from the analysis of variance (ANOVA) between the groups which results can be graphically illustrated by the boxplots of correct responses shown in Figure 5 and Figure 6, corresponding to the shorter announcements and longer announcements respectively. On the boxplots, each box has lines at the lower quartile, median and upper quartile values. The whiskers are lines extending from each end of

Table 2: Analysis of the responses from the activity of hearing and seeing and only hearing announcements.

	Announcement SA		Announcement SB		Announcement LA	Announcement LB
	Total # words (%)	Correct words (%)	Total # words (%)	Correct words (%)	Correct topics (%)	Correct topics (%)
Mean	81.82	71.43	72.73	63.64	54.55	77.27
Mode	100	100 and 71.43	60	100 and 40	100 and 33.33	75
Std. Dev.	23.12	26.34	24.12	32.02	30.81	17.52

the box to show the extent of the rest of the data. In the first case, for shorter announcements, Figure 5, the ANOVA shows that the audio and animation results cannot be considered statistically distinguishable since their statistical fluctuations can be confused. For longer announcements however,results from audio and facial animation can be considered statistically discernible with a statistic significance $p < 0.005$.
Other results were obtained from a 5-points Likert Scale questionnaire regarding technical aspects of the Virtual Presenter as speech and video quality.

Moreover, before the end of the workshop, a quick and informal final discussion reviewed their impressions on the activity.

The questions were either related to the voice that read the announcement (Festival), the video with a human face (Virtual Presenter), or related to the integration of both voice and video instantiated at *Vila na Rede*. These results are discussed in the next section.

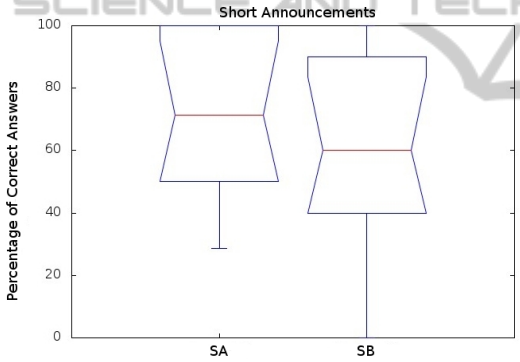

Figure 5: Boxplot of correct words for shorter announcements.

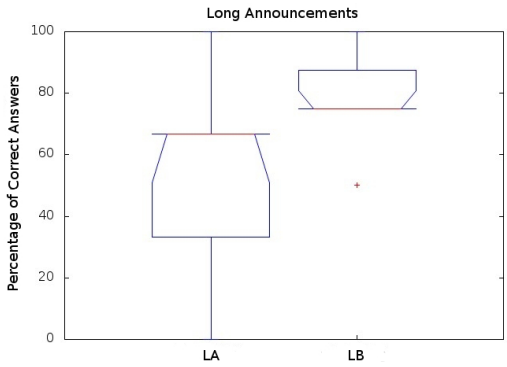

Figure 6: Boxplot of correct answers for longer announcements.

6 DISCUSSION

This Section presents a discussion regarding the results of the case study. Also, we expose some considerations involving the contribution of the approach to accessibility in ISN, technical and performance capabilities and limitations of using multimedia in context of socio-technical diversity, and other foreseen usage for the approach.

Analysis of Results. Based on the data from the responses of the activity of hearing and/or seeing announcements content we observed that: 1) the more precise transcriptions came from the audio-only media, for shorter announcements; 2) the more complete understandings came from using Virtual Presenter media, for longer announcements; 3) and there was a significant difference among the participants' individual results. In the transcription of the media content for shorter announcements, participants were slightly better when using audio-only content (71.43%) than using the Virtual Presenter (63.64%). We believe that the objective of the task contributed to this result; as the participants were really concerned about repeating the text they just heard, they might not pay attention to the video. Analyzing the statistical mode of the correct responses we observe that the virtual presenter obtained almost the same scores as the audio-only (*i.e.* 100% and 71.43% for audio-only, and 100% and 40% for Virtual Presenter).

When the activity consisted of writing down their understanding of the media content of the longer announcements, we observed a difference in the types of media. The Virtual Presenter obtained a rate of 77.27% of correctness while audio-only reached 54.55%. By analyzing the statistical mode we reinforce the difference *i.e.* a draw between 33.33% and 66.77% for audio-only, and 75% for Virtual Presenter. Additionally, the standard deviation of the Virtual Presenter is about a half (17.52%) of the audio-only, which indicates an actual improvement in the understanding of the media.

A comparative analysis between audio-only and Virtual Presenter for each participant results reveals that: results vary significantly among the participants (*e.g.* ranging from 19.64% to 95%). When considering the participants that obtained the best results, we identify that one got better results from the audio-only media and the other from the Virtual Presenter. However, when computing the results of the rate of correctness of the audio-only in relation to Virtual Presenter media for each

participant, we verified that 9 (out of 11) of them got better results using the Virtual Presenter.

In general, the acceptance of the voice alone was good. The results from the questionnaire confirmed our initial hypothesis that the participants enjoyed synthesized voices. Most of the participants reported to be able to clearly understand the voice and the opinion about its quality ranged from fair to good. But, as accounted by the questionnaire's results, the voice was perceived as mismatched with the video. The characteristics resulted from the combination of voice+video was reported to be unnatural and voice and video did not match (not in the sense that mouthing and speech were not encompassed, but that the voice did not seem to come from the person represented in the video). One factor that might also have impacted in the rating of the video is that most participants, as they were making intensive effort in the task of hearing and taking notes, did not pay much attention to the video. This was observed by researchers and also, it was added as a side note in the questionnaire, by one of the participants.

The data obtained from the observations and analysis of the forms (level of comprehension) matches the impressions they orally reported to have about the feature; *i.e.* that the video had a positive influence in the understanding of the spoken announcements.

Influence of Text Length. The contrast of the results in short and long announcements deserves some reflexion. For short announcements (5 to 7 words) it seems to be easier to keep all the announcement content in the short-term memory. So that a message using only audio was potentially better since participants were supposed to write the words, while a message using audio+video could request too much attention from the participant. On the other hand, for longer announcements, the focus was on understanding the information topics instead of only words. In that case, the message length was too big to be memorized, consequently, participants obtained better results paying more attention to the Virtual Presenter. Another aspect that could influence is the participants' literacy level. Despite of the fact that the activities consisted only of hearing and watching the messages in a controlled environment, it is necessary to have an investigation of literacy influence in the comprehension of spoken messages when using Virtual Presenter in the whole ISN, involving different subjects, contexts and writing formalisms.

Technical Performance. The preliminary evaluation conducted with the proposed TH during the experiment behaved as expected (*i.e.* without any technical problem). However, considering the social level of the target public, other important point to take into account regards the technical performance of the proposed solution. Since target users probably would not have Internet connection with large bandwidth in their house, it is important to try to predict the behaviour of the solution being used in such situations of restricted connection bandwidth. It is also relevant to note how much extra bandwidth the video consumes and the speed at which the information is delivered to users. Such technical performance details also would be useful in analyzing the benefits of the Virtual Presenter's usage (*e.g.*, in increasing understanding of the announcements using the solution in their home environment).

Accessibility. The concern with bandwidth and technical requirements is important, especially when aiming at more accessible applications. People should be able to profit from the feature, in despite of the technology or Internet connections available. This approach is in tune with the concept of ISNs, allowing users to access the information, regardless of the computer configuration available. The accessibility provided by the use of the Virtual Presenter goes further, making it possible for the functionally illiterate people, younger children, the elderly or anyone else who might have trouble with reading or seeing the letters to access to written information. Even some deaf communities should also be able to benefit from the resource, as the mouthing from the Virtual Presenter allows lip readers to understand the words (there are some communities of deaf people who are not taught how to read).

Foreseen Usage. The use of the Virtual Presenter is expected to contribute to the accessibility of ISN by reading its content, but without suppressing the users from the opportunity of having contact with the written information. Actually, it is expected that the reading of the text should contribute – even if in the smallest proportions – to alphabetization processes. The following of the text while read by someone else can be very useful also in other contexts, like learning another language. Other uses of the Virtual Presenter that could be foreseen include its use in e-books, e-mails, other web pages, information booths and ATM. Among the adaptations and developments needed for those uses, one of the most important would be to let the face express feelings and

intonation. Emotions in TH systems are already being studied (*e.g.* Cassell, 2001). In this sense, further investigation is needed to include emotion aspects in AnimaFace2D aiming at leveraging the Virtual Presenter's acceptance.

Other aspects to be considered in future investigations are: to check for results in more natural settings, *i.e.* situations where users are not demanded to provide written and accurate account from the audio, and to analyse the influence of individual skills regarding their educational and technological skills on the results obtained.

7 CONCLUSIONS

The Virtual Presenter a talking head that has been adjusted for our context has been incorporated into the Inclusive Social Network *Vila na Rede*, aiming at helping users to make better use of the information available in the system. A 2D facial animation that realistically reproduces speech articulatory movements was implemented at *Vila na Rede*, together with a female-converted voice from the Festival text-to-speech tool.

This paper described the process in which the Virtual Presenter was incorporated into our system and presents results of the activities that evaluated this mechanism. The results indicate that the video with a human presenter moving her lips accordingly might help users in the understanding of longer audio extracts. Nevertheless, it is important to have voice and face in harmony, providing a more natural aspect to the presenter.

The experiment contributed to clarify our target users preferences and use of this particular solution regarding a virtual presenter in the system; the results also suggest further studies and development regarding: tuning the solution towards more effective uses of a Virtual Presenter in inclusive systems scenarios, investigation of how people get appropriate of it in non controlled situations, and tests comparing the use of text-only and Virtual Presenter supported content.

ACKNOWLEDGEMENTS

This work is funded by Microsoft Research – FAPESP Institute for IT Research (#2007/54564-1), and partially by FAPESP (#2007/02161-0), CNPq (#383334/2008-0), and CAPES (#2001-8503/2008). The authors also thank colleagues from IC/UNICAMP, NIED/UNICAMP, InterHAD for insightful discussion.

REFERENCES

Brazilian Internet Steering Committee. 2009. *Survey on the use of Information And Communication Technologies in Brazil*. [online] Available at: www.cgi.br/english/index.htm [Accessed 21 January 2011].

Cassell, J., 2001. Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. In: *Embodied conversational agents.* Cambridge: MIT Press, pp. 1–27.

Cosatto, E., Östermann, J., Graf, H. P., Shroeter, J., 2003. Lifelike talking faces for interactive services. In: *Proceedings of the IEEE*, v. 91, n. 9, pp. 1406–1429. IEEE.

Costa, P. D. P., De Martino, J. M., 2010. Context-dependent visemes: a new approach to obtain realistic 2D facial animation from a reduced image database. In: *23rd SIBGRAPI Conference on Graphics, Patterns and Images - Workshop of Theses and Dissertations*.

E-Cidadania, 2011. [online] Available at: http://www.nied.unicamp.br/ecidadania [Accessed 18 January 2011].

FFmpeg, 2010. [online] Available at: http://www.ffmpeg.org/ [Accessed 15 November 2010].

Hayashi, E. C. S., Baranauskas, M. C. C., 2009. Communication and expression in social networks: Getting the "making common" from people. In: *Proceedings of the 2009 Latin American Web Congress, Joint LA-WEB/CLIHC Conference*, Mérida, Mexico. IEEE Computer Society (2009), pp. 131137.

Hayashi, E. C. S., Baranauskas, M. C. C., 2008. Facing the digital divide in a participatory way – an exploratory study. In: *Proceedings of the HCIS 2008, IFIP World Computer Congress (WCC 2008),* v. 272, pp. 143-154. Springer.

Institute Paulo Montenegro. 2009. Indicator of Functional Literacy. [online] Available at: <www.ipm.org.br> [Accessed 10 October 2010].

Johnson, L. W., Rickel, J. W., Lester, J. C., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, v. 11*, pp. 47–78.

Long Tail Ad Solutions, 2010. JW Player. [online] Available at: http://www.longtailvideo.com/players/jw-flv-player [Accessed 15 November 2010].

Mumford, E., 1964. *Living with a computer*. London, England: Institute of Personnel Management.

Neris, V. P. de A., Almeida, L. D. A., De Miranda, L. C., Hayashi, E. C. S., Baranauskas, M. C. C., 2009. Towards a Socially-constructed Meaning for Inclusive Social Network Systems. In: *11th International Conference on Informatics and Semiotics in Organisations*, IFIP WG8.1, pp. 247-254.

Open Computer Vision Library, 2010. [online] Available at: http://sourceforge.net/projects/opencvlibrary/ [Accessed 15 November 2010].

Östermann, J., Millen, D., 2000. Talking heads and synthetic speech: an architecture for supporting electronic commerce. In: *Multimedia and Expo, 2000. ICME 2000. IEEE International Conference*, v. 1, pp. 71–74. IEEE.

Pandzic, I. S., 2002. Facial animation framework for the web and mobile platforms. In*: Proceedings of the. Seventh international Conference on 3D Web* Technology (Web3D '02), pp. 2734. ACM.

Ryokai, K., Vaucelle, C., Cassell, J., 2003. Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of Computer Assisted Learning* v. 19, n. 2, pp. 195–208.

Schuler, D., Namioka, A., 1993. *Participatory design: principles and practice*s. Hillsdale, USA: Lawrence Erlbaum Assoc. Inc.

Stamper, R. K., Althans, K., and Backhouse, J., 1988. Measur: Method For Eliciting, Analysing and Specifying User Requirements. In: *Computerized Assistance During the Information Systems Life Cycle*. North-Holland: Elsevier Science Ltd., pp. 67-115.

Tartaro, A., Cassell, J., 2006. Using Virtual Peer Technology as an Intervention for Children with Autism. Universal Usability: *Designing Computer Interfaces for Diverse User Populations*. New York: John Wiley & Sons, Ltd., pp. 231–262.

University of Edinburgh, 2010. The Centre for Speech Technology Research - The Festival Speech Synthesis System. [online] Available at: http://www.cstr.ed.ac.uk/projects/festival/ [Accessed 15 November 2010].

Vila na Rede, 2011. [online] Available at´: http://www.vilanarede.org.br [Accessed 18 January 2011].