Usability evaluation of 3G multimodal services in Telefónica Móviles España

Juan José Rodríguez Soler(2),

Pedro Concejero Cerezo(2),

Carlos Lázaro Ávila (1),

Daniel Tapias Merino(1),

(1) Telefónica Móviles España
Serrano Galvache, 56, E-28033
Madrid (Spain)

(2) Telefónica Investigación y Desarrollo
Emilio Vargas, 6; E-28043
Madrid (Spain)

Abstract

Third generation (3G) services boost mobile multimodal interaction offering users richer communication alternatives for accessing different applications and information services. These 3G services provide more interaction alternatives as well as active learning possibilities than previous technologies but, at the same time, these facts increase the complexity of user interfaces. Therefore, usability in multimodal interfaces has become a key factor in the service design process.

In this paper we present the work done to evaluate the usability of automatic video services based on avatars with real potential users of a video-voice mail service. We describe the methodology, the tests carried out and the results and conclusions of the study. This study addresses UMTS/3G problems like the interface model, the voice-image synchronization and the user attention and memory. All the user tests have been carried out using a mobile device to take into account the constraints imposed by the screen size and the presentation and interaction limitations of a current mobile phone.

1. Introduction

Third Generation (3G) services boost mobile multimodal interaction offering users richer communication alternatives for accessing different applications and information services. These 3G services provide greater interactive flexibility and active learning possibilities but, on the other hand, may increase the complexity of the user interfaces.

Multi-Modal User Interfaces (MMUI) are particularly important in mobile services and applications because of:

(1) the limited attention and time that users have for interacting with services while they are in mobility,
(2) the broad range of different devices users have to handle with, and
(3) the large variety of use contexts.

Multimodal technologies have a key role to play in improving hand held devices and service usability, and also in creating new, compelling applications for mobile consumers.

In this outline, a major challenge at Telefónica Móviles España (TME) is to achieve MMUI which will minimize users' cognitive load and to provide robust MMUI, by avoiding interaction errors and providing active guidance and adaptation facilities to cope with different users' needs and demands.

1.1. Previous work in videotelephony

The video-call service is an example of one of TME 3G mobile video-telephony services which has benefited from usability evaluation. Video-telephony technology is not as new in the market as most people believe. It is actually a well-known service, from the ergonomic point of view, at least in fixed telephony [2][3][4]. It is surprising, then, that the evolution of video-telephony was not as booming as for example mobile telephony. This fact has not just been an issue of technical feasibility, but of fitness-for-purpose [5]. The coming of broadband and video-capable mobile phones has provided the basis to turn one-to-one real-time video communications from what has been a minor telecoms sector into a mass-market product. But usability evaluation still plays an important role in the successful video-telephony service mass-market launch.

At this point, after the positive experience of the video-call service, TME's strategic framework is to continue improving mobile services and terminals by means of usability evaluation in order to achieve a user-friendly approach.

Telefónica I+D, in collaboration with TME, has carried out studies to evaluate the usability of UMTS Video-telephony Services, and of Automatic Services based on Avatars and Automatic Speech Recognition (ASR) during spring 2004 and then during the year 2005, respectively. In the trials of the video-call service, part of the user sample included hearing-impaired users who communicated with sign language and the main results of the tests were included into a guide, on how to get the most from videotelephony, which is included in all the Movistar phones.

In both cases, research took place in TME's usability lab [6] before the service mass-market launch in order to implement users' feedback and expectations in the final version of the service and in the usability specifications of the 3G phones.

2. Method

2.1. Participants

The user tests were carried out with ten people, 5 men and 5 women, that were classified in two age groups: group 1, between 18 and 30 years old, and group 2, between 31 and 50 years old.

Previous experience about the use of voicemail service was also controlled: 30% of the participants were frequent users, 50% were sporadic users and 20% had never used the service.

2.2. Trials Material

There were several requirements for the generation of prototypes and the preparation of trial materials:

- Trials should be carried out on mobile terminals with QCIF resolution (176x144), because this is the typical resolution of most 3G phones.
- Prototypes should be able of appropriately playing video sequences with the avatar sound and with low frame rate (13-15 fps)

Several platform alternatives were considered, taking into account these requirements. In particular, up to 18 platforms were evaluated with the prototype contents, and the choice was finally for a PDA: the Hewlett-Packard IPAQ 6300.

All users could use the whole set of 18 prototypes, whose presentation order was previously randomized.

Last, all quantitative data was captured by means of three different questionnaires:

- A first questionnaire on socio-demographic information from the user.
- A second questionnaire for each prototype which captured the user judgments on each particular prototype.
- A final questionnaire to capture global evaluations about all the different aspects of the whole set of prototypes assessed.

2.3. Experimental design procedures

Experimental design combined the variables presented in table 1.

Variables	Levels	Labels
Interface model (IM)	4	**a**: Global IU
		b: "Sequence" IU
		c: Augmented IU #1
		d: Augmented IU #2
Synchronization Level (SL)	3	**a**: Synchronized
		b: Disarticulated
		c: Unsynchronized
The combination of DTMF options (TRIADA).	2	**a**: 8,4, 7 and 0 DTMF
		b: 6,3,5 and * DTMF
How do the options appear? (ORDER)	3	**a**: Three in screen
		b: first and second
		c: first and third

Table 1: Experimental variables

The experimental design used was a within-subjects factorial one, i.e., all participants used every combination of the above variables and levels, each one of them a different prototype.

IM	SL	TRIADA	Order	Prototypes
Global IU	a,b,c	a	a	3
"Sequence" IU	a,b,c	a	a	3
Augmented IU #1	a,b,c	a,b	a,b,c	6
Augmented IU #2	a,b,c	a,b	c,a,b	6
TOTAL				**18**

Table 2. Experimental design

Regarding the first experimental variable, interface model, trials included 4 different types. These were chosen to be different in both avatar size and in relative position. The presentation of options could be by means of icons or text labels, and the number of options available was also variable. In particular, the 4 interface types were:

Global IU

*The avatar appears in the **centre**, with **maximum size** and with **no other visual element** on display.*

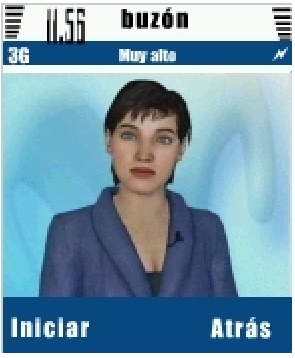

Figure 1.Global IU

"Sequence" IU

The avatar appears in the ***centre*** *of the screen and with* ***maximum size*** *but it disappears when the options are displayed.*

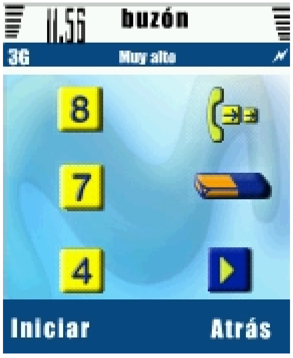

Figure 2. "Sequence" IU

Augmented IU #1

The avatar appears on the ***left side*** *of the screen and with* ***maximum size****, and, simultaneously, icons representing options appear on the right side with strict order, first the number of the option, and then, the icon.*

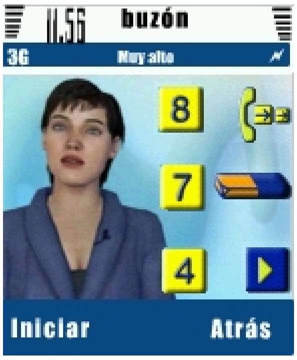

Figure 3. Augmented IU #1

Augmented IU #2

The avatar appears on the ***right side*** *of the screen and with* ***smaller size****, and, simultaneously, icons representing options appear on the right side with strict order, first the number of the option, and then, the icon.*

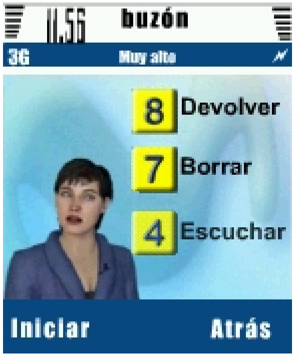

Figure 4. Augmented IU #2

The experimental variable synchronization was manipulated in two ways simultaneously: disarticulated and unsynchronised.

In addition, a control condition was developed in which avatar sequences were totally audio and video synchronized.

Disjoint (disarticulated) conditions

With this variable we intended to achieve a similar effect to that obtained with film dubbing. By means of this, the avatar lip articulation would not coincide with its associated speech tracks (phrase utterance) [9].

Several changes in the trial locutions viseme-phoneme were made with this purpose:

- Visemes corresponding to phonemes of open vowels were substituted by visually opposite visemes (bilabial-stops).
- The visemes corresponding to the close vowels were replaced with the visually opposite visemes (open vowels)
- The visemes corresponding to the bilabial-stop phonemes were replaced with the most open vowel (/a/)
- The visemes corresponding to the velar-stop phonemes were replaced with the close visemes (bilabial-nasals).
- The visemes corresponding to fricative phonemes were replaced with the close visemes (bilabial-stops).
- The visemes corresponding to vibrant phonemes were replaced with the visually opposite visemes (open vowels).

Unsynchronized conditions

These conditions were prepared to achieve an effect of time gap between the voice signal and the avatar lip movements, based on the literature recommendations about audio-video synchronization [7], [8].

Four levels of non-synchronization between audio and video were finally established, and were introduced in the trials:

- Slight audio anticipation with respect to video. The sequence audio begins 5 frames (200ms) before the original starting point.
- Severe audio anticipation with respect to video. The sequence audio begins 10 frames (400ms) before the original starting point.
- Slight video anticipation with respect to audio. The sequence locution begins 10 frames (400 ms) after the original starting point.
- Severe video anticipation with respect to audio. The sequence locution begins 20 frames (800 ms) after the original starting point.

The third experimental factor was called "triad", this refers to the group of three instructions which are presented to users so that they can interact with the service. For instance:

- To call back, introduce 8
- To listen to the next message, introduce 5
- To listen again to the next message, introduce 3

and so on.

Three measurements were made in the experiments:

1. Detection of the avatar synchronization variations in four different interaction moments with the trial service:

- Welcome to the service
- Received messages notification
- Reading of a message left by another user
- Information about the service handling options

2. DTMF commands recognition, as associated to the given instructions about the service handling. Two scores were used to obtain this measurement: the score of the number of correctly recognized keys and another score about wrongly associated keys.

3. Memory about service commands. Two scores were obtained in this case: first the number of correctly remembered service commands, and second, the number of wrongly remembered commands.

3. Results and Discussion

A series of one-way ANOVA was performed, with subsequent Scheffe pairwise multiple comparisons. All procedures used an alpha of 0.05.

3.1. Synchronization perception

First ANOVA with synchronization as experimental factor shows differences in the detection of synchronization variations, as a function of the evaluated moment. This difference appears in all interface models used in the experiments.

Please note: In this context, F means the ANOVA statistical contrast for differences between the means of all the conditions in the experimental factor. After the ANOVA, Scheffé pairwise multiple comparisons were carried out, and these contrasts are referred to as Scheffé F in the text.

More in particular, we can state that users detect the variations ONLY in the service welcome F(2,61)=6.965, $p<0.05$, when the user gets the received messages F(2,97)=13.002, $p<0.05$ and last, when they are informed of the available options F(2,34)=4.741, $p<0.05$.

When presenting the welcome to the service, users can perfectly detect if the avatar is synchronized or not (Scheffé F =1.48, $p<0.05$).

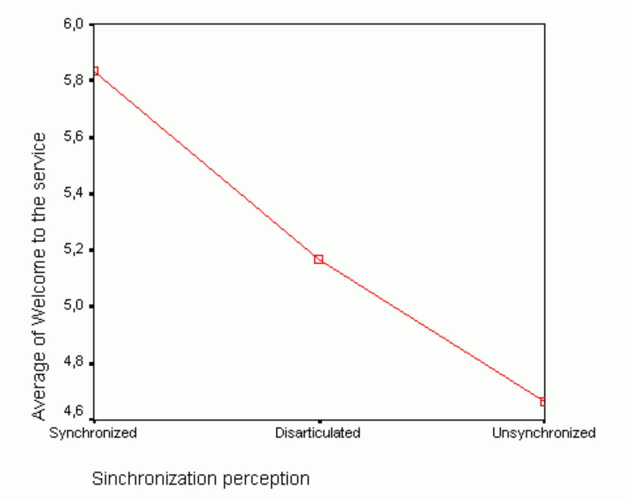

Figure 5. Means plot of synchronization perception of welcome to the service

When reporting about the reception of messages (number of messages and time), users can detect when the avatar is not synchronized Scheffé F=1.87, $p<0.05$, and also when it is disarticulated Scheffé F=1.85, $p<0.05$.

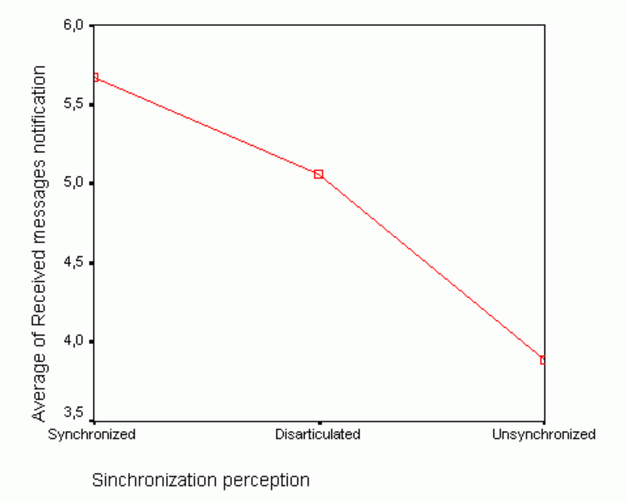

Figure 6. Means plot of synchronization perception in the received messages notification

And in the moment when the avatar is informing users about the service handling options, users can perfectly detect when the avatar is synchronized or not (Scheffé F=1.09, $p<0.05$).

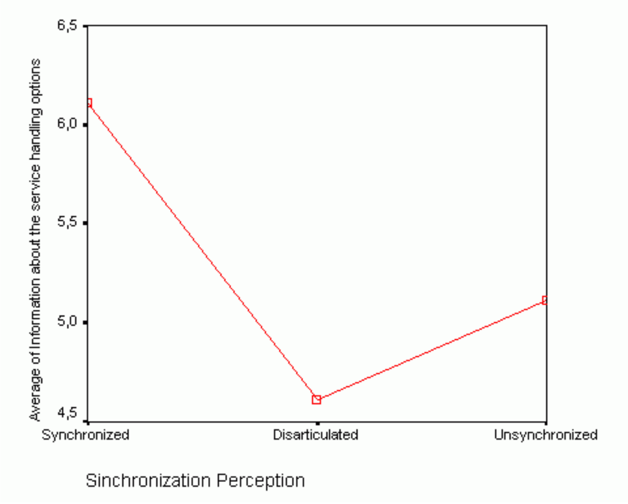

Figure 7. Means plot of synchronization perception in the information about the service handling options

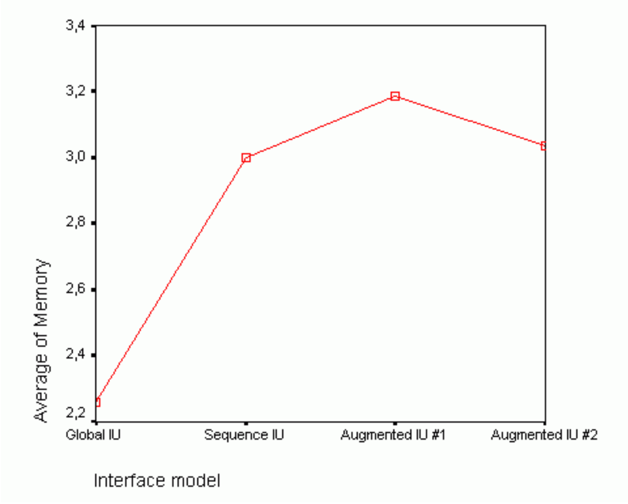

Figure 8. Means plot of number of correctly remembered options depending on the interface type

Therefore, the moment in which users detect more variations in avatar synchronization is when they are informed of the message reception.

The fact that this result does not appear in the moment of handling options presentation indicates that there are differential effects in synchronization perception depending on the specific interface model used.

3.2. The interface type makes impact in attention and memory

This chapter presents the results about the possible relationship between the user capabilities of visual recognition and the abilities to remember, with special focus on the multimodality effects.

First of all, we explored the possible relationship between the interface type and the recognition capability. There were no statistically significant differences F(3, 5)=2.288 , p>0.05.

However, we found statistically significant differences in the memory processes F(3, 16)=5.768 , p<0.05.

We can then confirm that the interface type makes an impact on the user capability to remember the service handling commands. More in detail, users remember in a different way when they judge the global interface with respect to the rest of interfaces (sequence, augmented#1 and augmented#2). However there are no differences in the memory associated to the evaluation to the later.

We can conclude that the best interface type is the augmented, i.e., the one in which the interaction is not made only with the avatar, but in different moments during the interaction with the service, the avatar is displayed simultaneously with the graphical information. For instance, presenting the numbers with the DTMF options associated with the handling commands.

All these results are closely related to the best known theoretical usability models, as that in ISO 9241:11, i.e., satisfaction, efficiency and efficacy.

- The perception of avatar synchronization makes impact in the user satisfaction with the service.
- The attention requirements (and consequently the cognitive load) is affected by the capabilities of recognition of the DTMF commands, and this is directly related with the interface efficiency.
- The memory capability to remember the service handling commands determines the efficacy of the interface.

4. References

[1] Linear Acoustic. (2004) Audio and Video Synchronization: Defining the problem and implementing solutions. *White Paper of Linear Acoustic, Inc.*, available online at: http://www.linearacoustic.com/wpapers.htm

[2] ETSI ETR 297 (1996). Human Factors (HF); Human Factors in videotelephony. Sophia- Antipolis, France: European Telecommunications Standards Institute.

[3] ETSI ETR 198 (1995). Human Factors (HF); User trials of user control procedures for integrated services digital network (ISDN) videotelephony. Sophia-Antipolis, France: European Telecommunications Standards Institute.

[4] ETSI ES 201 275 (1998). Human Factors (HF); User control procedures in basic call, point-to- point connections, for Integrated Services Digital Network (ISDN) videotelephony. Sophia-Antipolis, France: European Telecommunications Standards Institute.

[5] ETSI TR 102 274 (2003); Human Factors (HF); Guidelines for real-time person-to-person communication services. Sophia-Antipolis, France: European Telecommunications Standards Institute.

[6] Rodríguez, J. J.; Concejero, P. ; De Diego, S.; Collado, J. A.; Tapias, D.; Sánchez, A. J. (2005): Laboratorio de Usabilidad de Telefónica Móviles España (spanish text meaning "The usability lab at Telefónica Móviles España). Boletín de Factores Humanos, No. 27, available online at: http://www.tid.es/presencia/boletin/bole27/bol27_art03.htm.

[7] Linear Acoustic. (2004) Audio and Video Synchronization: Defining the problem and implementing solutions. *White Paper of Linear Acoustic, Inc.*, available online at: http://www.linearacoustic.com/wpapers.htm

[8] ITU-R BT.1359-1 (1998), Relative Timing of Sound and Vision for Broadcasting. International Telecommunications Union.

[9] Lewkowicz, D.J.(2000). The development of intersensory temporal perception: An epigenetic systems/limitations. *Psychological Bulletin*, 126,281-308.