Intl. Journal of Human–Computer Interaction, 30: 321–341, 2014

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2013.858461

Taylor & Francis
Taylor & Francis Group

An Empirical Investigation Into Text Input Methods for Interactive Digital Television Applications

Aurora Barrero, David Melendi, Xabiel G. Pañeda, Roberto García, and Sergio Cabrero
University of Oviedo, Asturies, Spain

Nowadays there is a huge market emerging in the interactive digital TV realm. In this context, we need new and effective methods of user interaction, as the main interaction device is still the classical remote control. Remote controls are especially problematic when it comes to writing text, something needed in most applications. Thus, we have carried out an empirical investigation to find effective methods of text entry with remote controls. We analyze several methods by performing experiments based on a methodology in which a heterogeneous set of real users carries out several sequential tasks in an incremental process. We analyze entry speeds, error rates, learning profiles, and subjective impressions, taking into account the particular characteristics of the users. Our results show, for instance, that Multitap is a good method for simple texts. It is between 12% and 34% faster than the fastest virtual keyboard, depending on the age of the user. Nevertheless, when complex texts need to be written, virtual keyboards present the same or even better writing speeds (QWERTY is 13% faster) and with significant lower error rates (Multitap is 347% worse than QWERTY). We consider that our results are very interesting for researchers, designers of TV applications, and hardware vendors.

1. INTRODUCTION

Recent innovations in the Internet protocol television (IPTV) field (Spira, 2011) and regulation changes (Congressional Record, 2009; European Union, 2005) have created a new scenario in the television realm in which content providers are encouraged to offer advanced services to be competitive. Most of these advanced services require interaction capabilities that are possible nowadays thanks to the proliferation of home Internet connections (Brandtzæg, Heim, & Karahasanović, 2011), the improvement of set-top-box and TV devices and the general evolution of telecommunication networks. The Internet and TV have merged, forming services such as IPTV or Internet TV (Spira, 2011). However, if complex applications are to be developed, we also have to design effective methods of user interaction, taking into account that the main interaction device available is still the remote control. This is a problem when it comes to writing text, as remote controls have not been designed for this purpose. Thus, if we need the users to introduce text on interactive digital TV (IDTV) applications, we have to find effective methods of text entry with a remote control considering aspects such as the experience of the users or their age (Taveira & Choi, 2009).

The main goal of this article is to provide a detailed study on text entry methods that may be used with commonly available technology that most people will probably already have in their homes. Thus, several methods are analyzed by performing experiments with a heterogeneous set of real users in an incremental process based on several sequential tasks. Some preliminary results published in Perrinet et al. (2011) are now improved with new results, more text entry methods, and usage contexts. We conclude answering questions such as which method is the fastest, which method is the one that leads to fewer user mistakes, how age affects performance, how performance improves with shortcuts or fast keys, how an extended set of characters diminishes performance, or how the layout of the remote control affects performance.

Our article presents an important set of contributions compared with previous work. We have analyzed a considerable number of input methods, some of them complemented with several improvements, such as disambiguation or suggestion mechanisms. We have also performed the evaluations under general contexts, also taking into account the special needs of other contexts and issues such as internationalization. Moreover, we have evaluated how the shape and the location of the keys in remote controls affect user performance. Finally, an important number of users participate in our study with a heterogeneous set of characteristics: age, gender, knowledge level, and technology usage habits.

The rest of the article is organized as follows. In Section 2 previous work is revised. The plan of the experiments is defined in Section 3. The evaluation and the obtained results are detailed in Section 4. Finally, conclusions and future work are presented in Section 5.

Address correspondence to David Melendi, University of Oviedo, 2.7.11 Edificio Polivalente, Campus de Viesques, s/n, Xixón, Asturies 33203, Spain. E-mail: melendi@uniovi.es

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

2. RELATED WORK

The study and comparison of different text input methods is a field where a lot of research has been carried out. Nevertheless, most recent advances have been focused on mobile devices such as smartphones or tablet devices. Although some of these advances may be applied to the IDTV realm, there are still many open issues as stated by Iatrino and Modeo (2006). First, the main interaction device is a remote control and not a keyboard, stylus, or touch screen. Nowadays this still holds true. Second, comparing TVs and mobile devices, the keys and the screen are in different devices, so users cannot look at the interaction device and the screen at the same time. Thus, interactive digital television services offer new challenges to be solved.

Commonly available remote controls can be used in only two possible ways: using cursors and OK keys over a virtual keyboard or using a reduced set of keys. Regarding virtual keyboards, in the field of mobile devices there have been several publications with interesting results for the IDTV realm. Bellman and MacKenzie (1998) compared two types of virtual keyboards. The first type is the well-known QWERTY layout, and the second is an optimized alternative named FOCL. FOCL proposes placing letters in the keyboard surrounding a predetermined letter depending on their probability of following this letter. Although the results show no significant differences between each method, FOCL can be considered as a feasible improvement for IDTV context. Zhai, Hunter, and Smith (2000) presented a new virtual keyboard layout called Metropolis, 40% faster than QWERTY, to improve writing performance in certain devices. The authors stated that existing layouts such as QWERTY have been designed to be used with both hands, so they are not appropriate for these devices. Brewbaker (2008) pursued the same goal but using a genetic algorithm to customize the size and language of the keyboard. The results of Brewbaker (2008) and Zhai et al. (2000) are similar, but their methods were not tested with real users. Although it is unknown whether these methods are good for real users, the algorithmic approach toward the design of a new layout can be interesting for IDTV applications.

Also, there are writing mechanisms for mobile devices based on a reduced set of keys suitable for IDTV applications. For instance, Silfverberg, MacKenzie, and Korhonen (2000) compared three methods: the multipress method used to write Short Message Service (SMS) messages (sometimes called Multitap) either with timeout or next key to switch between letters, a method based on two keystrokes called 2-key (first users select a group of letters and afterwards the desired letter), and the predictive method T9 (based on the theory that each combination of keystrokes produces a single word or a reduced set of words). These methods are evaluated with 12 users between 23 and 47 years of age, and the results show that the best results are obtained with T9 and the worst with 2-key. Similarly, Butts and Cockburn (2002) repeated the experiment of Silfverberg et al. (2000) but without the T9 method. Eight advanced users obtained the best results with the multipress method with next key and the worst with 2-key. MacKenzie, Kober, Smith, Jones, and Skepner (2001) presented a different disambiguation technique based on prefix probabilities called LetterWise. This method is compared with T9 in an experiment with 20 paid users, who participated in 20 sessions of 25 to 30 min. Users had to write sentences from a set defined by the authors. Whereas in the first session the results between methods were similar, in the last sessions there were some differences in performance. Nevertheless, they are unable to establish a significant difference between error rates for each of the methods. Nesbat (2003) presented a novel text entry system (MessagEase) for small electronic devices with a unique keyboard based on letter frequency and positional matrix. This keyboard is applicable to hard-key devices with a limited number of keys. The text entry system presented here provides full text entry (full ASCII 220) and is adaptable for any language. Based on the application of Fitts's law, this system is determined to be 67% faster than the QWERTY soft keyboard and 31% faster than multipress, but it is unknown whether it is good for users, as the results are only theoretical.

Apart from virtual keyboards and methods based on a reduced set of keys, there are other methods called "concurrent techniques" designed to combine different methods at the same time. Although there are several papers where these techniques are used, the results of Wigdor and Balakrishnan (2004) are especially interesting. They conducted a comparative experiment in which they add three additional keys to a conventional mobile phone keyset. Users have to combine these additional keys with regular keys depending if they want to write a number or a letter. They performed an experiment with 15 experienced users and detected a significant improvement in performance. These concurrent techniques can also be applied to an IDTV environment, for instance, when a greater set of symbols is needed.

The aforementioned papers mainly come from the field of mobile devices, but there are also a few noteworthy studies in which the issue of writing text in IDTV applications is treated. Iatrino and Modeo (2006) compared multipress (the SMS method), multipress with visual feedback, and a QWERTY virtual keyboard layout. Thirty-six people participated in an experiment where each user had to write an e-mail address and a short sentence in Italian. Results show that the best method is multipress. Also, the authors mention important problems with internationalization. Ingmarsson, Dinka, and Zhai (2004) presented a new technique called TNT, similar to TwoStick (Költringer, Isokoski, & Grechenig, 2007). Each character is accessible with two keystrokes with a similar approach to 2-key (Silfverberg et al., 2000). Five people between 27 and 32 years of age participated in the experiment during 10 sessions of 45 min to write a short novel in Swedish. Results show speeds comparable to or faster than manual writing in a personal digital assistant or the multipress method. The most valuable feature for the users of the experiment was the simplicity of the method. Geleijnse, Aliakseyeu, and Sarroukh (2009) compared Multitap

(multipress), T9, and a QWERTY virtual keyboard layout with the speed of a conventional QWERTY keyboard. The goal was to search for author-track pairs in YouTube. Twenty-two people between 21 and 32 years of age participated in the experiment. The authors conclude that the conventional keyboard is faster than remote control methods and that there are no significant differences between remote control methods, in contrast to previous works. Their final conclusion is also surprising: "User tests have not shown any indications that users do not accept the keyboard in a living room-like setting" p. 148. This conclusion contrasts with Orbist, Bernhaupt, and Tscheligi (2008), who consider that voice recognition systems may be a better solution. Their results show that it is not feasible to rely on the popularity of external peripherals for televisions different from those conventional. More recently, Gargi and Gossweiler (2010) presented a new predictive system designed to improve writing speed in virtual keyboards: QuickSuggest. This method shows a ring surrounding a given character, with the four characters that most probably follow it. If a user writes a symbol, he or she only has to select the next character with a cursor key and OK. Then the system moves automatically to the written character and it starts over again. The authors perform a theoretical performance study and an experiment with 10 real users. Sporka, Polacek, and Slavik (2012) compared TNT (Ingmarsson et al., 2004) and a new method named TwiceTap. This method has the same philosophy as TNT but, apart from allowing users to type single characters, it also allows them to write frequent blocks of characters (n-grams). Eighteen paid users, with an average age of 22.7, participated in the experiment. There were no significant differences between the methods, but users preferred TwiceTap mostly because of its similarity to Multitap and the availability of n-grams.

2.1. Discussion and Problem Description

As mentioned previously, there is little research on text writing methods for IDTV applications if we consider that the main interaction device is still the remote control.

In recent years there have been other papers, apart from those mentioned previously, aiming to solve the problem of writing text. Nevertheless these studies rely on interaction devices that have nothing to do with a remote control. This is the case of Wobbrock, Myers, and Aung (2004) and Költringer et al. (2007) studying joystick text entry methods; Oniszczak and MacKenzie (2004) combining keystrokes with finger movements in RollPad; Orbist et al. (2008) and Vega-Oliveros, Pedrosa, Pimentel, and De Mattos Fortes (2010) with voice recognition; MacKenzie, Lopez, and Castelluci (2009) and Aoki, Maeda, Watanabe, Kobayashi, and Abe (2010) with pointing devices; Rick (2010) and Varcholik, LaViola, and Hughes (2012) with mechanisms of writing text on interactive surfaces; or Choi, Han, Lee, Lee, and Lee (2011) with an ad-hoc-designed remote control equipped with a touchpad. Although the results of all these previous works are very interesting and may be applied to IDTV environments in the future, they are not eligible with current massively available TV technologies, as the main interaction device is still the conventional remote control.

As we have seen, there are only a few studies about text entry in TV environments with methods suitable for remote controls. Also, there are even fewer studies with results obtained from real users. Many papers present conclusions based on Fitts's prediction models (MacKenzie, 1991) or other theoretical analyses. Their performance with real users is unknown. Other papers perform evaluations with a reduced set of users or with very homogeneous characteristics: usually young people with a technical background. Other papers study the same or similar methods but present contradictory results. Furthermore, there are important issues that have not been studied in depth, such as the impact of a complex set of symbols, the shape of the remote control, or the location of the keys. We consider that the issue of writing text on IDTV applications with a remote control needs further study.

In our opinion, our article presents new results regarding text input issues in the IDTV realm. First, we have analysed several types of virtual keyboard layouts and other methods, some of them complemented with different improvements, such as disambiguation techniques or suggestion systems. Second, we have performed the evaluations using both general sentences and specific characters and texts. Third, we have evaluated how the shape and the location of the keys in remote controls affect user performance. The last important characteristic of our study is the number of users and the variations in their characteristics: age, gender, level of knowledge, and technology usage habits.

3. TEST PLAN

3.1. Goals and Metrics

The main goal of the experiments was to compare several input methods for IDTV applications. To achieve this goal, various metrics were initially chosen:

- Measured entry speed: in characters per minute. We considered the time taken to write a certain sentence, including the time spent in writing the desired sentence, but also the time spent in deleting mistaken characters and rewriting the correct ones.
- Measured error rate: percentage of mistaken characters written. To calculate error rates we counted the total amount of characters written by the user and compared it with the number of characters in the desired sentence.
- Measured learning curve: improvement (or not) of users' performance.
- Subjective impression of ease of use (0 = *very difficult* to 4 = *very easy*).
- Subjective impression of speed of use (0 = *very slow* to 4 = *very fast*).
- Subjective impression of user satisfaction (0 = *unsatisfied* to 4 = *very satisfied*).

We have focused the analyses on entry speed and error rates because if we observe previous work, we can clearly see that all the authors use them as basic metrics to compare text entry methods. Nevertheless, as shown in MacKenzie et al. (2001), the values of these basic metrics evolve when users participate in several sessions in the experiments. Thus, it is also important to study how entry speed and error rates vary with experience.

Apart from empirical data, it is also important to collect subjective information. Previous work shows that the opinions of the users are very valuable and that these do not always coincide with empirical performance data. For instance, Sporka et al. (2012) showed no empirical differences between methods but a clear preference of the users toward TwiceTap. We have chosen speed of use to compare the impression of the users with the measured speed. We have also selected user satisfaction as a good indicator of the general impression of the users, as in Iatrino and Modeo (2006). Finally, we have chosen ease of use because this is one of the most valuable features for users in previous work (Ingmarsson et al., 2004).

3.2. Subjects

In total, 96 users participated in the experiments. These participants were as heterogeneous as possible, with ages ranging between 20 and 64, different technological skills, gender, and laterality. The participants did not receive any money or any other kind of compensation. Some of their characteristics are shown in Table 1.

Not all the users participated in all the tasks, so further details are provided in the description of each task. We grouped users according to their age as follows: users belonging to the "mobile phone generation," ages 18 to 30 (young); users belonging to the "computer generation," ages 31 to 45 (adult); and "pre-pc" users, older than 45 (older). We have focused the analyses on age because changes in perceptual and motor skill capabilities that accompany the aging process bring implications for the design of human–computer interfaces (Taveira & Choi, 2009). This is aligned with the differences reported by Siek, Rogers, and Connelly (2005). The rest of the data in Table 1 are descriptive. Nevertheless, as we expect users to behave differently depending on their attitude toward technology, we have taken into account the habits of the users using the information gathered in several questionnaires. Finally, we have also considered gender and laterality, because these are also common types of groups of demographic segments in usability.

3.3. Apparatus

Hardware. During the experiment, each participant was left alone in a room, to avoid distractions. To create the most realistic situation possible, users sat as if they were in their own living room. An armchair was placed in front of a 32-in. LCD television of dimensions 814 × 599 mm placed on top of a table around 1 m in height. According to the recommendations of the television manufacturer, the armchair was placed at a distance of 2 m (the optimal distance varies from 1.5 to 3 m for this type of display).

The television was connected to a PC with Microsoft Windows XP running an application designed to carry out the tests. We also connected several remote controls to the PC equipped with USB RF or infrared receivers, which incorporated the most common keys of conventional television remote controls. This decision was based on the fact that one of our design goals was to investigate text entry methods that may be used with commonly available technology that most people would probably already have in their homes. In most of the tests we used a SnapStream Firefly PC remote control, shown in Figure 1.

Software. We developed an Adobe AIR application to carry out the tests. This application, shown in Figure 1, had a black full-screen layout and one or more text input fields to allow the users to write. Our fonts were Verdana 65 point for the proposed sentences and the input fields, and Arial 60 points for the keyboard layouts. Depending on the test, the text to be written was placed on top of these input fields.

The application was developed to process keyboard events. To map remote control strokes into keyboard events, we used the EventGhost software. No forced time lags between keystrokes were introduced by the system, and users were allowed to delete symbols.

The application showed a 10-s counter at the beginning of the experiment and between different input methods. This counter was also used before writing each sentence, allowing users to

TABLE 1
Characteristics of the Participants

		Gender		Field (Profession or Studies)		Education		
Age	Total	M	F	IT	Non-IT	Elementary	High School	University
Young	57	32	25	39	18	0	7	50
Adult	27	17	10	14	13	0	5	22
Elderly	12	8	4	1	11	5	1	6

Note. IT = information technology.

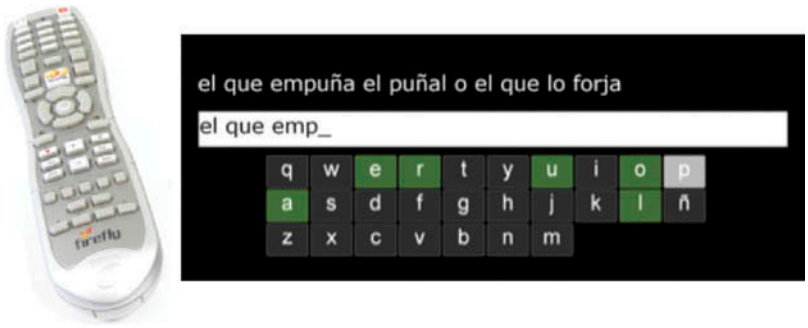

FIG. 1. Remote control and testbed application.

read sentences in advance. Once the user was allowed to write, the application checked whether the sentence was written completely. When a sentence had been completed, the application switched to a new sentence or to a different input method.

When virtual keyboards were used, the keyboard layout was placed on the bottom part of the screen, and users had to use the cursors and the OK button to move around the layout. We also included the following improving mechanisms:

- Users could move faster by keeping a cursor button pressed. This reduced their physical load. Based on pilot tests, we adjusted both the start of the auto-repeat function and its repetition rate to 0.2 s.
- The borders of the layout were wrapped around both vertically and horizontally. For instance, users could go from the top part to the bottom by pressing the up cursor and vice versa.
- We implemented a suggestion system inspired by the LetterWise disambiguation method (MacKenzie et al., 2001) and similar to those in FOCL (Bellman & MacKenzie, 1998) and QuickSuggest (Gargi & Gossweiler, 2010). Based on what a user wrote, the system changed the color of the six letters most likely to follow what was already written, as shown in Figure 1. To provide these recommendations, the system loaded a dictionary of the Spanish language (RAE – Royal Spanish Academy, 2001), and each time a user added a letter to a word, the system searched for all the words containing the string the user had written. With this list of words, the system computed the most frequent letters after the current string and displayed the six most probable with a green background.

When mobile-like techniques were used, the bottom part of the screen was used to show help information, as in Figure 2. The keys in the application were placed in accordance with the European standard ETSI ES 202 130 V2.1.2 (2007-09). To assist users when they pressed the same key several times, the current letter was highlighted.

The application stored information of the activity of the users in XML log files. In these files the application registered when a certain key in the remote control was pressed, information about

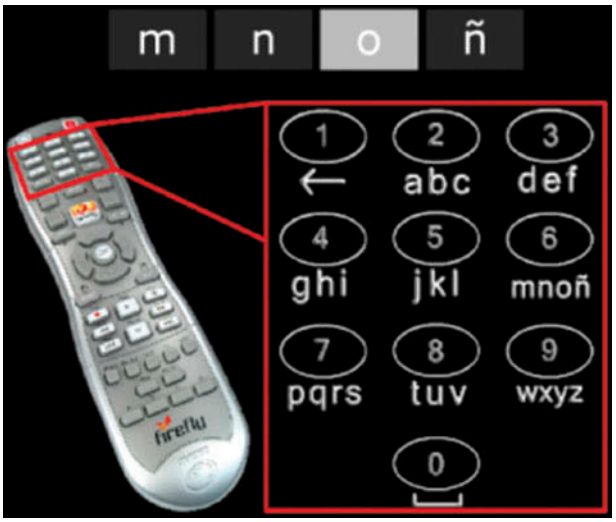

FIG. 2. Implementation of a mobile-like method in our application.

the answers to questionnaires and information about both the sentences proposed by the system and the sentences written by users. Further details of the format of these files may be found in Perrinet et al. (2011).

3.4. Text Input Methods

During the experiment, we compared the performance of several text input methods suitable for interactive TV applications. These methods can be classified into two groups: virtual keyboards and mobile-like methods.

- Virtual keyboards:
  - QWERTY: This is the traditional keyboard layout, broadly used in previous work.
  - Square alphabetic—or simply alphabetic: This is based on the alphabet, the position of the keys in alphabetical order.
  - Genetic: We used a genetic algorithm to generate a new keyboard layout in order to improve the efficiency of the users. Using a modern version of Don

Quixote as a reference text for Spanish, the algorithm positioned the keys so that the most used letters appear in the center of the keyboard. The resulting layout is shown in Figure 3. The details and cost model of this algorithm can be found in Brewbaker (2008).

- Modified QWERTY: This was QWERTY with an additional row of symbols including special vowels as shown in Figure 4. Also, four additional buttons in the remote control were used. One allowed the user to switch to a layout with capital letters, another allowed the user to switch to another layout with special symbols (Figure 5), and two more buttons allowed the user to write blank spaces and delete.
- Modified Genetic: Using the genetic algorithm we generated a new layout but including special vowels. The result is shown in Figure 6. Also, four additional buttons in the remote control were used as in the modified QWERTY layout, with exactly the same purpose.

FIG. 3. Genetic keyboard layout.

- Mobile-like methods:
  - Multitap: Based on the mobile phone system, this method uses number keys to write text as shown in Figure 2. To differentiate between multiple strokes corresponding to a single letter and strokes between successive letters, we use a threshold of 1 s as in a regular phone.
  - T9: This is a predictive system based on the idea of pressing one single key for each letter of the word the user wants to write (Silfverberg et al., 2000). For example, if we consider the relation between the keys and the letters in Figure 2, the word "this" can be written by pressing 8, 4, 4, and 7. In most cases, the word that appears after pressing a sequence of keys is the desired one. When there is more than one possibility matching the sequence of keystrokes, users must select one of the available options.
  - 2-key: This is a system designed to reach any letter with only two keystrokes (Silfverberg et al., 2000). Using numeric keys, a first keystroke is used to select a group of letters and a second keystroke is used to select the desired letter.
  - Modified Multitap: In this version of the Multitap method, new characters were inserted in the keyboard with special vowels as shown in Figure 7. Also, we used a concurrent technique to change the functionality of the keyboard as in Wigdor and Balakrishnan (2004). A button allowed the user to switch to capital letters and numbers. Another button allowed the user to insert special characters.

FIG. 4. Modified QWERTY layout.

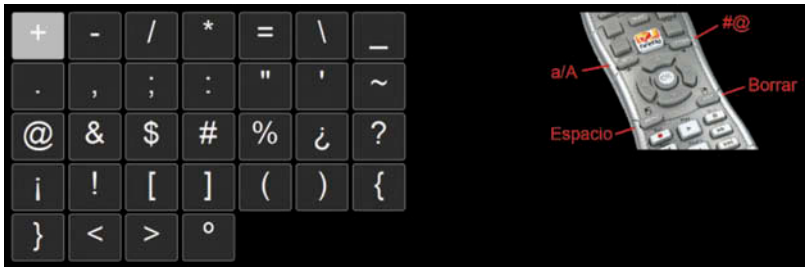

FIG. 5. Special characters layout.

FIG. 6. Modified genetic layout.

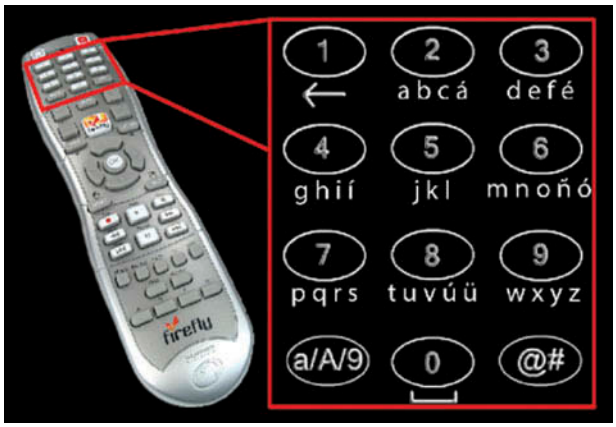

FIG. 7. Modified multitap layout.

Regarding virtual keyboard layouts, we have chosen QWERTY because it is already known by a great number of people familiar with computers and because it is the main layout used in previous work. This method was designed placing many adjacent letter pairs (digraphs) on the opposite sides of the keyboard, facilitating the frequent alternation of the left and right hand. But, on a virtual keyboard, the polarizing common digraphs mean that the user has to move back and forth more frequently and over greater distances than necessary (Zhai et al., 2000). Nevertheless, it is still currently the most popular layout. The Alphabetic layout was considered because it is supposed to be the "easiest" in the sense that anyone who knows the alphabet would be able to use it without a learning process. Thus, nonexpert users would find it very easy to use according to Zhai et al. (2000). We have used the Genetic layout, which had never been used in Spanish, as a more experimental option and obtained very promising results. We conducted a theoretical analysis in order to check the efficiency of this layout. Measuring the minimum number of keystrokes needed to write Don Quixote, this method is 45% faster than QWERTY for a proficient user. Finally, we have used the modified versions of QWERTY and the Genetic layouts to include symbols that are nowadays common in applications and symbols required in languages different to English.

On the other hand, mobile-like methods were chosen mainly inspired by Silfverberg et al. (2000). We consider that Multitap is interesting for the IDTV domain. Although it is progressively disappearing as modern mobile phones incorporate physical keyboards or touch screens, Multitap was extremely popular among young users not so long ago. Thus, the number of expert users of this method is still huge. This situation combined with the fact that previous works show that it is more efficient than virtual keyboards (Iatrino & Modeo, 2006) are the reasons why we considered this method for the evaluation. Also, T9 is a promising method according to the results in Silfverberg et al. (2000), surpassing the results of Multitap. We have chosen 2-key because, in theory, it is faster than Multitap for Spanish if we compare its two keystrokes with a calculated weighted average of 2.23 Multitap keystrokes per letter for the words in the official Spanish dictionary (RAE – Royal Spanish Academy, 2001). Finally, as in the case of virtual keyboards, we have a version of Multitap with an enhanced number of symbols to be used in real IDTV applications and languages different to English.

3.5. General Procedure

The experiment consisted of four tasks, each of them requiring the users to attend one or several sessions in which the same text entry methods were used. A summary of the plan of the experiment is shown in Table 2, and a brief description of each of the tasks follows:

- In the first task, we started by testing popular methods of text entry with a predefined set of sentences: QWERTY, Alphabetic, Genetic, and Multitap.
- In the second task we included several improvements, both proposed by users during the first task or gathered from previous works such as Silfverberg et al. (2000). The methods were QWERTY, Genetic, T9, and 2-key.
- In the third task we performed tests with personalized sentences (not predefined) and nonalphanumeric characters, using the modified versions of QWERTY, Genetic, and Multitap. Previous works such as Ingmarsson et al. (2004) or Iatrino and Modeo

TABLE 2
Summary of the Plan of the Experiment

Task	Users	Methods	Remote Controls	Sessions	Set of Characters	Texts
1	57	QWERTY, Alphabetic, Genetic, Multitap	SnapStream Firefly	5	Basic	5 sentences of corpus
2	48	QWERTY, Genetic, T9, 2-key	SnapStream Firefly	2	Basic	5 sentences of corpus
3	42	Modified QWERTY, Modified Genetic, Modified Multitap	SnapStream Firefly	1	Complex	5 fields with personal data and 1 common text
4	42	Modified QWERTY, Modified Genetic, Modified Multitap	Golden Interstar, AverMedia RMKS	1	Complex	5 fields with personal data and 1 common text

(2006) reported internationalization problems and, partially, problems with special symbols (@, commas, etc.). We considered this task important due to these problems and bearing in mind the deployment of IDTV applications in contexts such as web browsing or e-commerce.
- In the final task we performed tests with different remote controls to see whether the device itself influences performance. We carried out this task due to the lack of standard remote control layouts (Nielsen, 2012), as only general recommendations exist (European Broadcasting Union, 2006).

All these tasks were not designed a priori. When we first performed the experiments published in Perrinet et al. (2011), we discovered other problems that needed to be taken into account. First, the impact of certain optimizations, analysed in the second task. Second, the usage of special symbols to accommodate specific contexts, analyzed in the third task. Third, the influence of the design of the remote control, analyzed in the final task. These problems were further analysed in the aforementioned tasks, conforming the full experiment presented here.

As shown in Table 2, in some of these tasks we planned several sessions to measure how users learn. We tried to confirm the relation between experience and performance (MacKenzie et al., 2001). During the first session of each task, we gave the participants a brief explanation and an example of each input method. No instructions were given regarding how to handle the remote control, that is, with the left hand, with the right hand, loosely, tightly, and so on. Thus, we expected them to use the remote as they usually do. The methods were assigned to participants in a counterbalanced order to neutralize learning effects (analyses of variance [ANOVAs] for test order showed no significant differences).

In the first and second tasks we provided users with the texts to write. We created a small corpus with 50 representative short sentences in Spanish gathered from local newspapers. The sentences had an average of five words, and they included only numbers and lowercase letters from the English alphabet, with the single exception of the ñ symbol.

We used several questionnaires to gather users' details and subjective information:

- During their first session, users had to fill in a questionnaire with details of their gender, age, profession, level and type of studies, habits of using television, mobile phone and computer, and whether they were left- or right-handed.
- In all of the sessions users had to complete a final questionnaire providing us with some feedback. They had to comment on each method, for example, if they had the impression they were improving, and rate three Likert scales (0–4): ease of use, entry speed, and global satisfaction.

Finally, to extract the conclusions of the experiment we performed several statistical analyses. The main goal was to determine if there were real differences between the means of two or more groups of variables. We used the most common tests for this type of analysis, according to Crawley (2007). These tests depended on the assumptions of normality and homoscedasticity of the data. First, we checked normality with Shapiro–Wilk tests and homoscedasticity with Bartlett tests. When data met both normality and homoscedasticity, we used one-way ANOVA tests to compare the data. If homoscedasticity failed, then we used Kruskal–Wallis tests. Kruskal–Wallis tests were also used in situations with a strong failure in normality (p values in normality tests over .05). Finally, if differences existed, we used Tukey tests with a confidence coefficient of 95% to perform pairwise comparisons.

4. EVALUATION OF TEXT INPUT METHODS

4.1. First Task: Evaluation of Main Input Methods

In this task we compared the performance of four text entry methods suitable for IDTV applications: QWERTY, Alphabetic

and Genetic virtual keyboards, and Multitap. As previously stated, these methods were chosen after an analysis of previous work.

To track the progression of the participants, users had to complete five sessions on 5 consecutive days. In each of these sessions, they had to write five random sentences of the corpus with all the methods. The particular characteristics of the participants are shown in Table 3.

Some preliminary results of this task were published in Perrinet et al. (2011) with fewer users than in the present study.

Analysis of writing speed and learning profiles. The general results show that the fastest method is Multitap, whereas the slowest is QWERTY. The average writing speed in characters per minute for each method and Sessions 1 to 5 is shown in Figure 8. It is noticeable that Multitap is much faster than the rest in all sessions. Also, performance regularly improves with experience for all the methods. We have calculated how this progression may be in future sessions using the power law of learning (Ritter & Schooler, 2002) with the results shown in Figure 8. In this figure we can see a great progression in Multitap. For virtual keyboards the Genetic layout presents a slightly better tendency with similar results for QWERTY and Alphabetic. Moreover, our initial assumption was that QWERTY does not require a learning period, but learning does in fact occur as in the case of Clarkson, Clawson, Lyons, and Starner (2005).

To check whether the differences between each method have any statistical significance, we have performed pairwise comparisons with Tukey tests. Basically, Tukey applies simultaneously to the set of all pairwise comparisons $\{\mu i\text{-}\mu j\}$. Confidence intervals including 0 are not significantly different, and all the other pairs are significantly different. Confidence intervals greater than 0 represent greater values in the i set than in the j set, whereas confidence intervals lower than 0 mean the opposite. These comparisons present differences between the methods in all the sessions, but they are statistically significant only in the case of Multitap ($p < .001$), which is the fastest method. This coincides with Iatrino and Modeo (2006), but is in clear contrast with Geleijnse et al. (2009). For example, Figure 9 shows the results of these comparisons for Session 5. Genetic is the best virtual keyboard, QWERTY

TABLE 3
Characteristics of the Participants in the First Task

		Gender		Field (Profession or Studies)		Education		
Age	Total	M	F	IT	Non IT	Elementary	High School	University
Young	27	15	12	20	7	0	3	24
Adult	19	12	7	12	7	0	2	17
Older	11	7	4	0	11	5	1	5

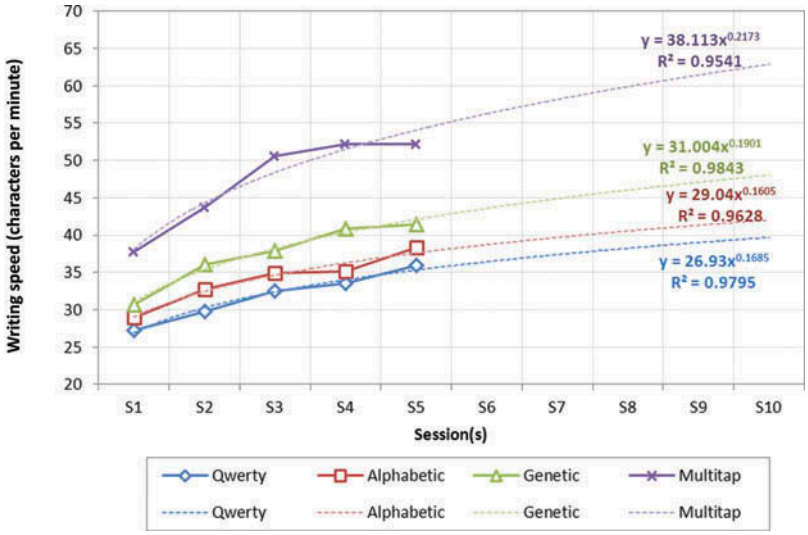

FIG. 8. Average writing speed per method and session and prediction of improvement with further experience.

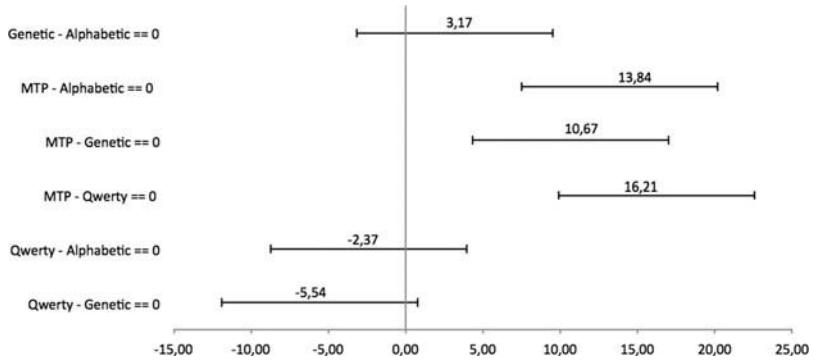

FIG. 9. Pairwise comparisons of average writing speed in Session 5 (95% family-wise confidence level). *Note.* MTP = Multitap.

being the worst, but this cannot be considered statistically significant.

If we compare empirical data with the habits reported by the users, we see that users who reported sending more than one SMS message per day are faster than users who send messages less frequently, but only for Multitap. The average speed is 58.87 characters per minute for frequent SMS users versus 46.60 for the rest of the users. Also, the participants who use computers frequently are faster in all the methods. Users with more than 12 hr a week of computer usage show statistically significant differences with the rest of the users in all the virtual keyboards ($p < .02$). In the case of Multitap, this difference is statistically significant only if we compare users with more than 12 hr a week with users with less than 3 hr a week ($p = .007$). Although users with weekly computer usage between 3 and 12 hr are faster in general than users with less than 3 hr, this is not statistically significant according to Tukey tests. Habits reported by users regarding their usage of TV do not seem to affect performance.

In the results we have seen that age is a major factor affecting performance, as shown in Figure 10. In general, the younger the user, the faster he or she writes. Pairwise comparisons of average speeds show that young users and adult users are faster than older users with a statistically significant difference ($p < .001$). Also, although young users are faster than adult users, this difference is statistically significant only for Multitap ($p = .0043$) and Alphabetic ($p = .02$).

Finally, ANOVA tests show us that there are no statistically significant differences between male and female users and left versus right-handed users.

Analysis of error rates and learning profiles. The general results show that the text input method with the highest error rate is Multitap, in contrast with virtual keyboards that present much lower error rates as shown in Figure 11. Multitap is worse than the rest in all the sessions. Moreover, in general for all the methods, error rates decrease with experience. Using the power law of learning we have calculated how this progression may be in future sessions with the results shown in Figure 11. In this figure we can see a great progression in Multitap and similar results for all the virtual keyboards.

Pairwise comparisons performed with Tukey tests show that only the differences between Multitap and the virtual keyboards are statistically significant making the Multitap the worst method considering error rates. For instance, in Session 5 we obtain a $p <$ 1e-05 if we compare Multitap ($M = 7.26$, $SD = 5.59$) with Genetic ($M = 1.68$, $SD = 1.37$), QWERTY ($M = 2.11$, $SD = 2.27$), and Alphabetic ($M = 1.67$, $SD = 1.72$). In general, Genetic is the best method but with results similar to those of QWERTY and Alphabetic.

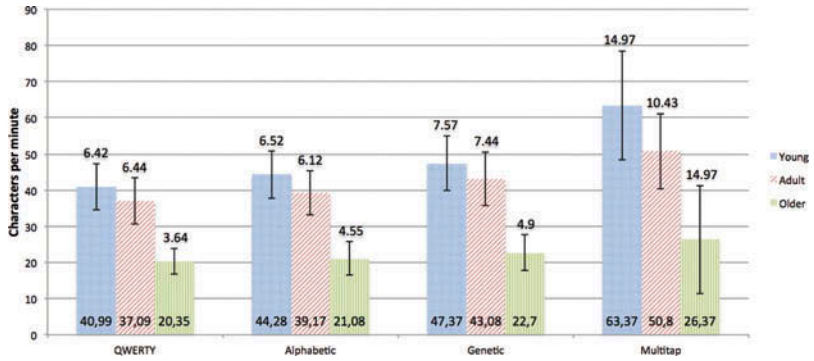

FIG. 10. Mean and standard deviation values of writing speed per method and age.

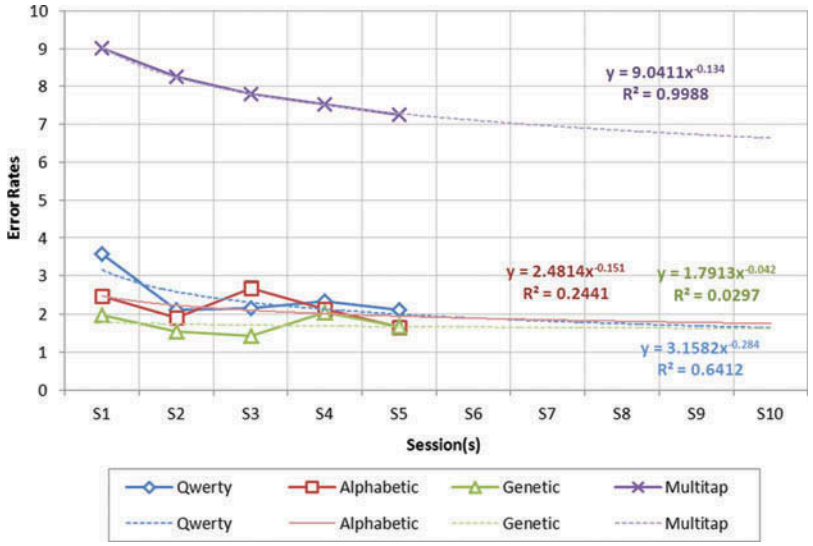

FIG. 11. Average error rates per method and session and prediction of improvement with experience.

In the results we have seen that age also affects performance. Pairwise comparisons of average error rates show that older users commit more errors than young and adult users with a statistically significant difference for all the methods ($p < .04$) but for the Alphabetic. There are no significant differences between young and adult users. Figure 12 shows 95% family-wise confidence intervals for the methods with statistical significant differences.

If we compare the empirical data with the habits reported by the users, we see that there are no differences in error rates for users with different habits toward SMS or TV services. Nevertheless, error rates depend to some extent on computer usage habits but only for QWERTY and Genetic. In general, the higher the computer usage, the lower the error rates are for these methods. This is clearer for users with more than 12 hr a week of computer usage, who present lower error rates than the rest of the users with statistically significant differences ($p < .047$).

Finally, the results present no differences between male and female users and left- versus right-handed users.

Subjective impressions. As mentioned previously, users had to fill in questionnaires in each session reporting their degree of satisfaction and their impressions about speed and ease of use. Figure 13 shows the mean scores of these values in Sessions 1 and 5. In the figure we can see how users improve their scores as sessions pass by. Not only do they feel faster with experience (with the single exception of Alphabetic), which coincides with empirical data, but they also feel more confident in general with all the methods.

If we compare subjective scores with empirical data, we can see that slight differences exist. First, if we use empirical data to rank the methods according to their speed, we obtain in all the sessions the same classification, as shown in Figure 8. On the other hand, if we rank the methods according to subjective impressions, we see how users change their minds as the experiment evolves. For instance, users feel that the fastest method in the first session is Multitap, which coincides with empirical data, but in the last sessions they feel that the fastest method is the Genetic. Second, the subjective impressions do not coincide

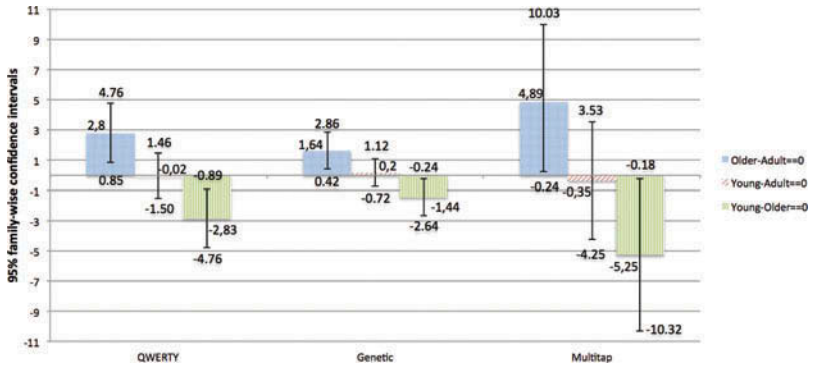

FIG. 12. 95% family-wise confidence intervals for error rates in methods with significant differences.

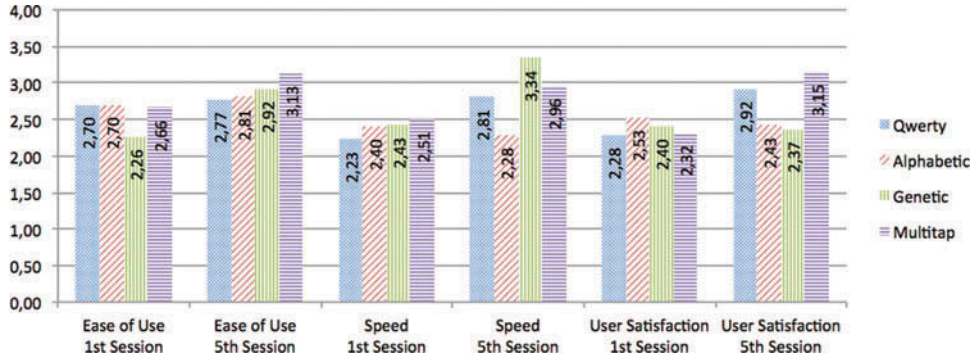

FIG. 13. Mean scores of subjective impressions per method in Sessions 1 and 5 (0 = *unsatisfied* to 4 = *very satisfied*).

with empirical data. For example, at the end of the experiment users found that the slowest method was the Alphabetic and the fastest the Genetic, which contrasts with empirical data that shows Multitap the fastest and QWERTY the slowest. Third, although empirical data reflect a good performance of the Genetic method, this is the worst method according to user satisfaction at the end of the experiment. It is considered to be easy and fast, but users were not satisfied perhaps due to the confusing position of the letters. Finally and surprisingly, users considered the Genetic to be the fastest method, but they were much more satisfied with Multitap, even though this is the worst method according to empirical error rates.

4.2. Second Task: Evaluation of Optimized Methods

In this task, only three of the four methods used in the previous phase were used: QWERTY, Genetic, and Multitap. Alphabetic was discarded because it did not show any significant differences with QWERTY in the previous task. The methods were modified to include a set of optimizations. In the previous task, many users highlighted that in virtual keyboards they had to cross the layout to insert a blank space or to delete a symbol. They suggested using keys of the remote for these actions, and in this task we tried to check whether this improves performance. Also, two enhancements of Multitap taken from previous work were considered: T9 and 2-key (Silfverberg et al., 2000).

Each user carried out two sessions on 2 consecutive days. With regard to the previous task, we reduced the number of sessions, as we were more interested in the effect of the improvements than in the analysis of learning patterns. In every session, users had to write five random sentences of the corpus with all the methods.

The particular characteristics of the participants are shown in Table 4. Some of the users who participated in this task also participated in the first task. Twelve of these users were in the group young, seven in adult, and four in older. This situation will be taken into account, as their previous experience may affect the results.

Analysis of writing speed. If we compare the performance of all the methods, pairwise comparisons performed with Tukey tests present clear differences between them. These comparisons are statistically significant ($p < .001$) in the case of T9, which is the fastest method ($M = 58.59$, $SD = 18.81$). Also, comparisons are statistically significant in the case of 2-key ($p < .05$), which is the slowest method ($M = 32.34$, $SD = 8.39$). Genetic ($M = 39.84$, $SD = 8.34$) is slightly faster than QWERTY ($M = 38.39$, $SD = 6.75$), but the difference is not statistically significant. Comparing these results with those gathered in the first task, we are able to detect whether the optimizations have been effective. To reach more interesting conclusions, we have different remarks for users who participated in the first task and users who did not. The results are shown in Table 5. For repeating users, we have compared the

TABLE 4
Characteristics of the Participants in the Second Task

		Gender		Field (Profession or Studies)		Education		
Age	Total	M	F	IT	Non IT	Elementary	High School	University
Young	33	21	12	26	7	0	3	30
Adult	10	7	3	7	3	0	2	8
Older	5	3	2	1	4	2	1	2

TABLE 5
Differences Between the First Task and the Second Task

	Repeating Users		New Users	
	Task 1 Session 5	Task 2 Session 2	Task 1 Session 2	Task 2 Session 2
QWERTY	38.56	38.05	29.74	38.62
Genetic	44.34	39.43	36.06	40.10
Multitap – T9	52.60	56.13	43.71	60.20
Multitap – 2-key	52.60	33.04	43.71	31.88

Note. Units: characters per minute.

speed in the last session of each task. For new users, we have compared the speed of the users of the first task in the second session of that task with the speed of new users in the second task in their second session, as all of them had the same degree of experience.

For repeating users pairwise comparisons show that there are no statistically significant differences for QWERTY, Genetic, and T9. Nevertheless, the results obtained with 2-key are 37.17 % worse than with the standard Multitap, with a statistically significant difference. In the case of new users, they improve performance with the optimizations for QWERTY ($p < .001$), Genetic, and T9 ($p < .001$), although in the case of Genetic the difference is not statistically significant. Again, the performance with 2-key is worse than with the standard Multitap ($p < .001$). In conclusion, the results show that new users improve faster with the optimizations than without them for all the methods but the 2-key. Also, a little experience greatly improves performance, reaching similar results within a reduced period of time. It is also interesting to see that the results of Multitap, T9, and 2-key methods coincide with those in Silfverberg et al. (2000) and in Butts and Cockburn (2002), but they are in contrast with those in Ingmarsson et al. (2004; TNT having a two-keystroke approach similar to 2-key) and Geleijnse et al. (2009).

If we compare the empirical data with the habits reported by the users, although there are some differences in speed depending on the usage of the SMS service, computers, TV, or T9 in mobile phones, these differences are not statistically significant. Mean values are shown in Table 6. It is very interesting to see that the results obtained with the T9 method by frequent T9 users ($M = 65.86$ characters per minute) are not particularly better than those obtained by other users ($M = 57.08$ characters per minute).

In the results we have seen that age is a major factor affecting performance, as shown in Figure 14. Again, the general criterion is that the younger the user, the faster he or she writes. Pairwise comparisons of average speeds show that young users and adult users are faster than older users with a statistically significant difference in all the methods ($p < .023$) but the 2-key. In the case of the 2-key, the difference between adult and older users is not significant. Also, although young users are faster than adult users, this difference is statistically significant only for T9 ($p < .015$) and 2-key ($p < .01$).

Finally, ANOVA tests show us that there are no statistically significant differences between male and female users and left- versus right-handed users.

Analysis of error rates. In general, virtual keyboards show much lower error rates than the other methods. Tukey tests show that T9 is the worst method ($M = 20.92$, $SD = 18.7$), with statistically significant differences with the rest of the methods ($p < .001$). Genetic is the best method ($M = 1.63$, $SD = 1.5$) with a significant difference with 2-key ($M = 7.41$, $SD = 7.06$; $p = .028$) but not with QWERTY ($M = 3.2$, $SD = 2.31$). Finally, the difference between QWERTY and 2-key is not significant.

We can compare these results with those of the first task, with different conclusions for users who participated in the first task and users who did not (Table 7). For repeating users, we

TABLE 6
Writing Speed Depending on User Habits

	QWERTY	Genetic	T9	2-Key
SMS messages per day	<1: 38.62	<1: 40.11	<1: 58.58	<1: 33.60
	1–5: 35.97	1–5: 37.58	1–5: 55.03	1–5: 28.47
	>5: 40.74	>5: 42.91	>5: 59.68	>5: 39.69
Hours of computer per week	<3: 38.46	<3: 40.69	<3: 57.51	<3: 41.15
	3–12: 36.32	3–12: 36.69	3–12: 60.35	3–12: 32.23
	>12: 38.05	>12: 39.72	>12: 57.20	>12: 31.89
Hours of TV per week	<7: 37.91	<7: 38.96	<7: 57.72	<7: 30.77
	7–21: 38.18	7–21: 40.40	7–21: 57.64	7–21: 33.44
	>21: 34.12	>21: 30.80	>21: 54.20	>21: 28.87
Usage of T9	No: 37.93	No: 39.66	No: 57.08	No: 31.15
	Yes: 36.71	Yes: 36.94	Yes: 65.86	Yes: 35.26

Note. Units: characters per minute. SMS = Short Text Messages.

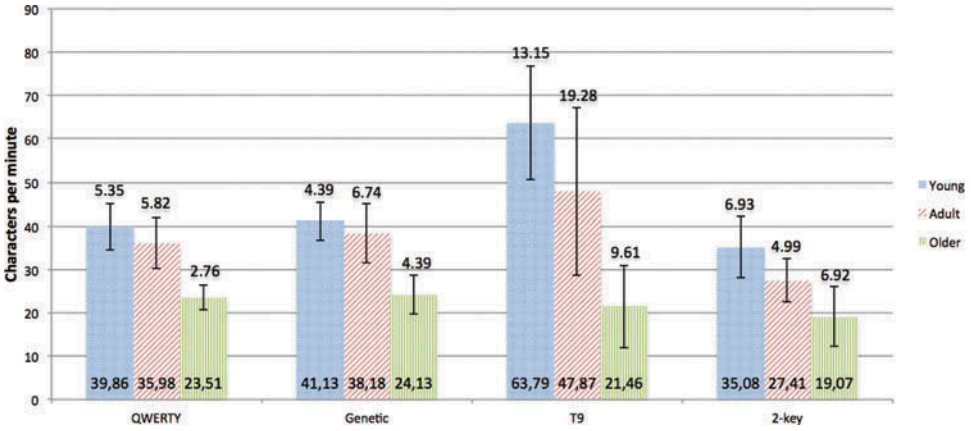

FIG. 14. Mean and standard deviation values of writing speed per method and age.

TABLE 7
Differences in Error Rates Between the First Task and the Second Task

	Repeating Users		New Users	
	Task 1 Session 5	Task 2 Session 2	Task 1 Session 2	Task 2 Session 2
QWERTY	2.05	2.91	2.11	3.39
Genetic	1.49	1.94	1.55	1.50
Multitap – T9	8.04	24.47	8.27	18.60
Multitap – 2-key	8.04	7.58	8.27	7.31

have compared error rates in the last session. For new users, we have compared error rates in the second session, as all of them had the same degree of experience. Pairwise comparisons show that the only statistically significant difference for repeating users is with T9, with an increase of 204% in error rates. Furthermore, for new users the difference is statistically significant for QWERTY ($p < .001$), with an increase of 60.18%, and T9 ($p < .001$), with an increase of 124.87%.

Although there are slight differences in error rates depending on the age of the users or their habits, pairwise comparisons show that these are not statistically significant. Also, there are no differences between male and female users and left- versus right-handed users.

4.3. Third Task: Evaluation of Methods With Special Characters and Specific Information

In the previous tasks we considered only numbers and lowercase letters. However, if applications need to be developed in a global environment, new symbols are needed. First, there are symbols needed to solve internationalization issues as happened in Iatrino and Modeo (2006) or Ingmarsson et al. (2004). Second, certain contexts such as web browsing, messaging, e-commerce, or e-government require special symbols (@, slashes, commas). Thus, tildes, accents and special symbols need to be included somehow if we would like to work under realistic conditions. To accommodate these situations, the input methods chosen for this task are the modified versions of QWERTY, Genetic, and Multitap.

To ensure the use of these special characters, the users had to fill in a form with their personal information including name, ID number, date of birth, address, and e-mail. All the users also had to write a common piece of text: an Internet internationalized resource identifier (this IRI is the same for all the users). An example follows:

- Name and surnames: Xurde Rodríguez Fonticiella
- Personal identification code: 12345678X
- Date of birth: 20/02/2002
- Home address: C/ My Address, n°11
- E-mail address: myemail.address@email.com
- Internet IRI: http://es.wikipedia.org/wiki/Teclado_informática

As the system did not provide the texts to be written, we followed a different approach to calculate entry speed and error rates. Entry speed was calculated once the form was completed by using the number of characters in the form and the total time taken by the user. To calculate error rates we counted the characters deleted by the users.

This experiment took place on a single day, during a session of about 35 min. We performed only one session, as the goal was to analyze the increase in the complexity of the methods. The particular characteristics of the participants are shown in Table 8. Although some of the users who participated in this task also participated in the first and/or second tasks, we have not considered this situation in the analyses because the design of this task differs greatly from the design of the previous tasks. Thus, we consider any previous experience irrelevant for the results of this task.

Analysis of writing speed. In this case, the fastest method is QWERTY (M= 30.66, SD = 6.71) followed by Multitap (M = 27.07, SD = 6.92), and Genetic (M = 26.97, SD = 6.49).

TABLE 8
Characteristics of the Participants in the Third and Fourth Tasks

		Gender		Field (Profession or Studies)		Education		
Age	Total	M	F	IT	Non IT	Elementary	High School	University
Young	24	17	7	20	4	0	2	22
Adult	13	9	4	9	4	0	1	12
Older	5	3	2	1	4	2	1	2

Nevertheless, an ANOVA test shows that there are no significant differences between the methods ($p = .058$). If we compare these results with those gathered in previous tasks, we can see that they are much worse. If we compare the results of virtual keyboards in Tasks 2 and 3 (optimizations are used in both tasks) with QWERTY, the speed decreases 20.15%, whereas with Genetic the decrease is 32.30%. Moreover, if we compare the results of Multitap in Tasks 1 and 3 (T9 and 2-key are not used), speed is now 48.09% worse. Clearly, there is an increase of complexity in usability that has a great impact on performance.

If we compare the empirical data with the habits reported by the users, the differences in speed depending on the usage of SMS, computers or TV are not statistically significant. Mean values are shown in Table 9. The only exception is Multitap: Users who send more than five SMS messages per day are faster than the rest of the users ($p < .045$).

Again, we have seen that age affects performance. As in previous tasks, the younger the user, the faster he or she writes. Pairwise comparisons show that young and adult users are faster than older users with a statistically significant difference in all the methods ($p < .01$). Furthermore, the difference between young and adult users for Multitap is also significant ($p = .018$), which coincides with the results of previous tasks.

TABLE 9
Writing Speed Depending on User Habits

	QWERTY	Genetic	Multitap
SMS messages per day	<1: 30.30 1–5: 28.61 >5: 37.86	<1: 27.28 1–5: 26.17 >5: 30.39	<1: 27.38 1–5: 23.81 >5: 36.97
Hours of computer per week	≤12: 30.28 >12: 30.39	≤12: 25.85 >12: 27.24	≤12: 26.34 >12: 26.93
Hours of TV per week	≤7: 28.26 >7: 31.50	≤7: 26.35 >7: 27.56	≤7: 23.44 >7: 28.71

Note. Units: characters per minute.

We have not found any statistically significant differences between male and female users and left- versus right-handed users.

Analysis of error rates. As in previous tasks, Multitap produces higher error rates than virtual keyboards. Tukey tests show that Multitap is the worst method ($M = 16.86$, $SD = 9.71$), with statistically significant differences with other methods ($p < .001$). QWERTY ($M = 4.85$, $SD = 3.51$) is slightly better than Genetic ($M = 5.49$, $SD = 5.65$), but the differences are not significant.

If we compare these results with those gathered in previous tasks, we can see that error rates are much worse. If we compare the results of virtual keyboards in Tasks 2 and 3 with QWERTY error rates, these increase 48%. With Genetic they increase 211%. Also, if we compare the results of Multitap in Tasks 1 and 3, error rates increase 163%. Again, the increase of complexity in usability has a great impact on performance.

Although there are slight differences in error rates depending on the age of the users, pairwise comparisons show that these are not statistically significant. Also, we have not found any significant differences if we compare the methods with users' habits. The only exception is Genetic: Users who use computers more than 12 hr a week have lower error rates than users who do not (4.67% vs. 12.85%). Finally, the results also show that there are no differences between male and female users and left- versus right-handed users.

4.4. Fourth Task: Evaluation of Different Remote Controls

In this case, users had to use remote controls with different keysets. As pointed out by Nielsen (2012), the differences in the design of remote controls due to the lack of any standardization constitute a huge usability problem. Thus, the goal of this task is to discover if changes in the shape of the remote control or the location of keys affect user performance. In this experiment two additional remote controls were used, as shown in Figure 15.

The remote control on the left of the figure had been used in the rest of the tasks. It is a SnapStream Firefly remote control, as mentioned previously, and is referred to as "remote #1." We chose the remote control in the center of the image because the number and arrow sets are placed to one side. The numbers are

FIG. 15. Remote controls used in the fourth task.

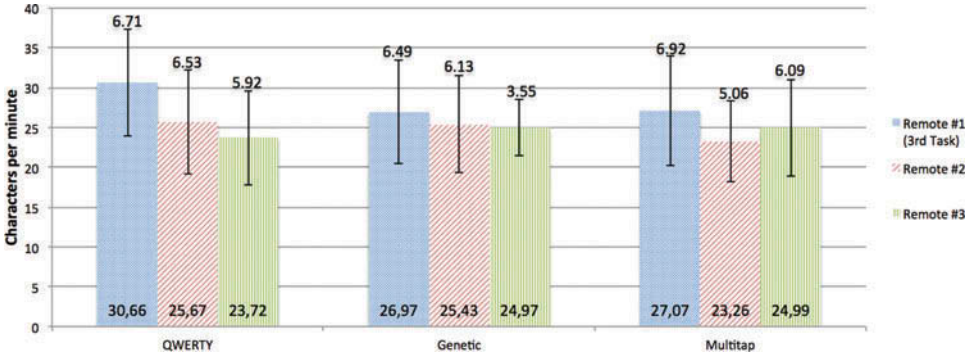

FIG. 16. Mean and standard deviation values of writing speed per method.

situated on the top-left part and the arrows and OK keys on the center-right part. It is a Golden Interstar remote control and will be referred to as "remote #2." We chose the remote control on the right of the image to check whether the size was important or not, as it is notably smaller than the rest. Also, this remote allowed us to check the impact of reversing the order of its elements, as the numbers are on the bottom-part and the arrows and OK keys on the top part. This is an AverMedia RMKS remote control and is referred to as "remote #3."

To compare the results with those gathered in previous tasks, we repeated the third task but using the new remote controls. This allows us to reuse the data gathered in the third task by comparing it with the data obtained with remote controls #2 and #3. Obviously, this time the form had to be filled in twice, each time with each of the new remote controls. To make data comparable, we used exactly the same group of people in the third and fourth tasks, so their particular characteristics are shown in Table 8.

Analysis of writing speed. In this case,the results are slightly worse than in the third task, as shown in Figure 16. If we consider independently the results of each remote control, there are differences in speed between each of the methods but they are not statistically significant according to ANOVA tests ($p > .05$). The same happened in the third task with remote #1.

If we perform pairwise comparisons using all the information gathered in the third and fourth tasks, we obtain the best and worst pair remote-method. As shown in Figure 17, the best results are obtained with remote #1 when QWERTY is used, with statistically significant differences in some cases (p ranging from less than 0.001 to 0.04 in those cases). On the other hand, the worst results are obtained with remote #2 and Multitap, but the differences are not significant according to Tukey tests.

Also, in the same pairwise comparison we can see whether the shape or location of the keys affect performance. For this purpose, we can analyze each method independently by comparing its pairs with all the remote controls (e.g., compare all the pairs of QWERTY with each other). In Figure 17 we can see that the only statistically significant difference is with remotes #1 and #3 with QWERTY ($p = .003$). Thus, our results show

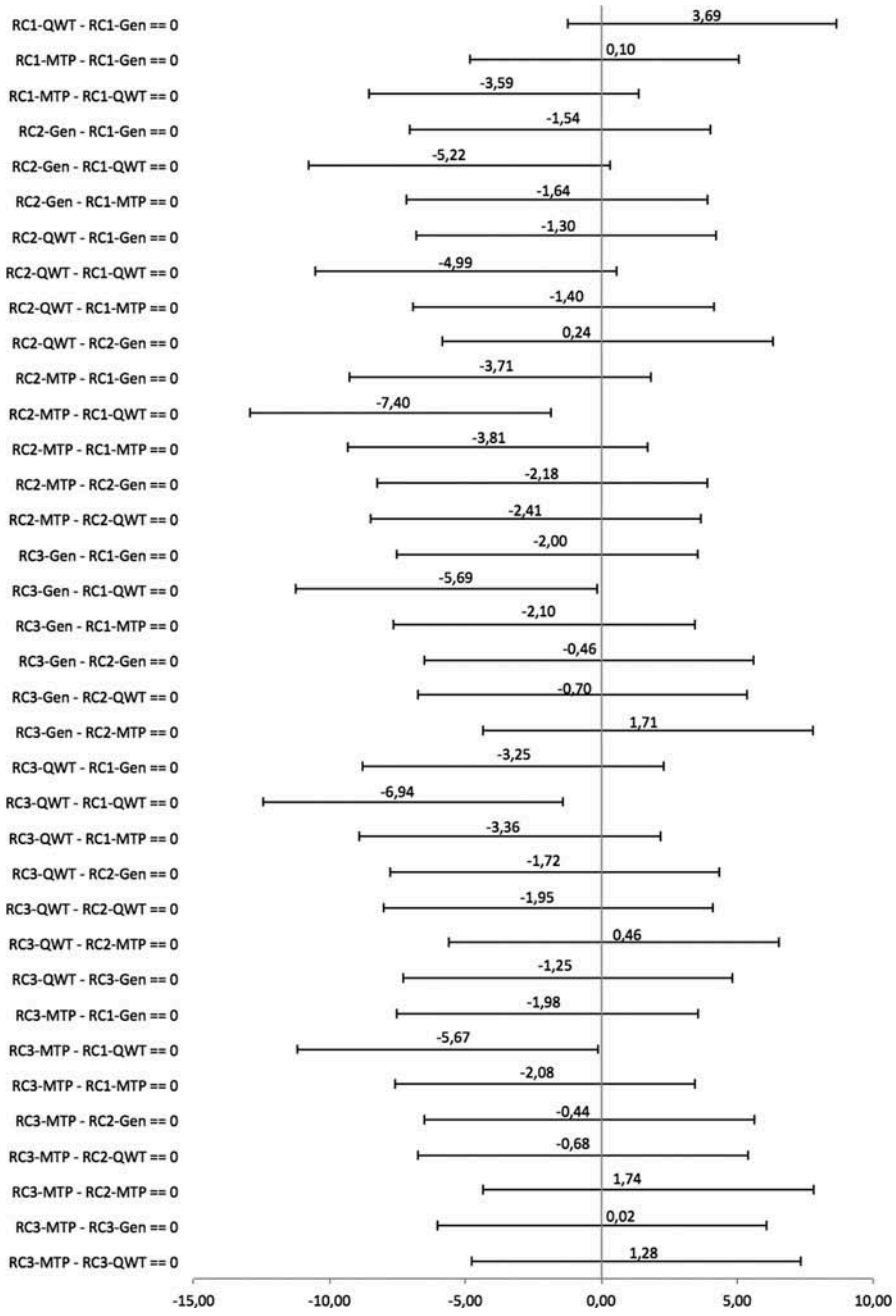

FIG. 17. Pairwise comparisons of writing speeds for each method and remote control (95% family-wise confidence level). *Note.* MTP = Multitap.

that, in general, the shape and location of the keys does not produce significant differences in speed.

Analysis of error rates. As in previous tasks, error rates registered with Multitap are much higher than with the rest of the methods in every remote control, as shown in Figure 18. Furthermore, the difference between Multitap and the rest of the methods is statistically significant in remote controls #2 and #3 as shown in Figure 19 ($p < .001$). The same happened in the third task with remote control #1.

As with writing speed, if we perform pairwise comparisons using all the data of the third and fourth tasks, we obtain the best and worst pair remote-method. As shown in Figure 19, the best results are obtained with Genetic and remote #2. The differences are statistically significant with Multitap for all the remote controls, as stated previously. On the other hand, the worst results are obtained with remote #1 and Multitap with statistically significant differences with virtual keyboards for all the remote controls according to Tukey tests.

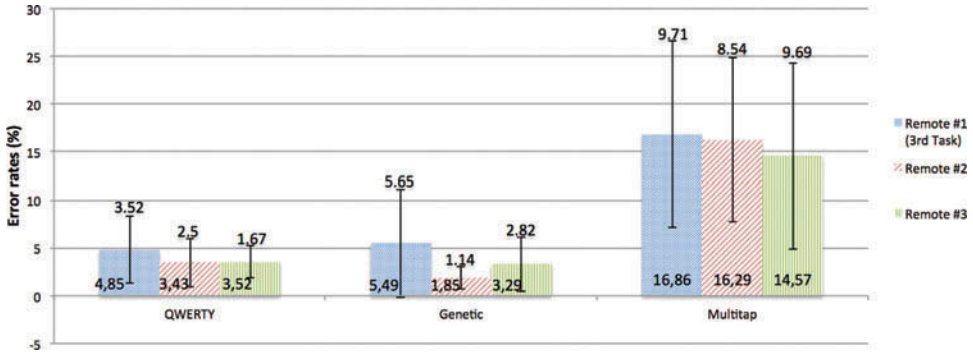

FIG. 18. Mean and standard deviation values of error rates per method.

As with writing speed, we can see if the shape or location of the keys affects error rates in the same pairwise comparison. In Figure 19 we can see that there are no significant differences if we analyze each method independently. Our results show that, in general, the shape and location of the keys does not produce significant differences in error rates.

5. CONCLUSIONS AND FUTURE WORK

Our article deals with human interaction in the IDTV realm with an important set of contributions. We have analyzed several types of virtual keyboard layouts and other methods, some of them complemented with different improvements, such as disambiguation techniques or suggestion systems. We have performed the evaluations using both general sentences and specific characters and texts, attending to issues such as internationalization or usage in particular contexts. We have evaluated how the shape and the location of the keys in remote controls affect user performance. Also, we present results gathered with a considerable number of real users with heterogeneous characteristics: age, gender, level of knowledge, and technology usage habits. These contributions are especially important if we consider that there is little research on text writing methods for IDTV applications with remote controls.

Our results show that, when simple texts need to be written, the method with the best performance is Multitap, despite the fact that this method produces higher error rates than virtual keyboards. Nevertheless, when writing complex texts, virtual keyboards present the same or even better writing speeds than Multitap and with significantly lower error rates. Using concurrent techniques has been fairly positive as stated by Wigdor and Balakrishnan (2004). When we combine virtual keyboards with certain buttons in the remote control for frequent actions we improve performance considerably. A logical conclusion is that users improve with experience, but we have seen that each method evolves differently: Multitap and the Genetic virtual layout show the best progressions. Another interesting conclusion is that subjective impressions do not always coincide with empirical results, and we think that the opinion of users can be as important or more than real performance. Furthermore, our results show that the shape and location of the keys in remote controls does not significantly affect performance. Another general conclusion is that user habits may have an impact on performance in some cases, but not on a general basis. Although performance is not affected by certain user characteristics such as gender or hand orientation, age is a major factor affecting performance: the younger the user, the faster he or she is. Performance tends to be similar between mobile-like methods and virtual keyboards for older users. Another interesting conclusion is that certain optimizations are not as efficient as they were intended to be, which is confirmed by the results obtained with methods such as T9, 2-key or Alphabetic virtual keyboard layouts.

Some recommendations for developers of IDTV applications follow:

- It is a good idea to provide multiple methods so users can choose the method they are most comfortable with. Performance is important, but we have seen that subjective opinions may indicate something totally different.
- If only one method can be used, there are better and worse methods depending on the context of the application. If simple tests need to be written, Multitap or T9 are good options. On the other hand, if special symbols are required virtual keyboards are better, as a wider range of users would be satisfied. In the latter case, the Genetic is a good option as performance improves greatly with experience.
- Use concurrent techniques to take advantage of the rest of the buttons of the remote control. For instance, spare buttons can be used for frequent actions such as deleting characters or writing blank spaces.
- Apparently, the design of the remote control does not have an impact on performance, but remote controls should be designed with enough distance between buttons to avoid fat-finger problems (Siek et al., 2005).

One of our design goals was to investigate text entry methods that may be used with commonly available technology that most people will probably already have in their homes. Thus, future work will be mainly based on experiments with devices different to conventional remote controls.

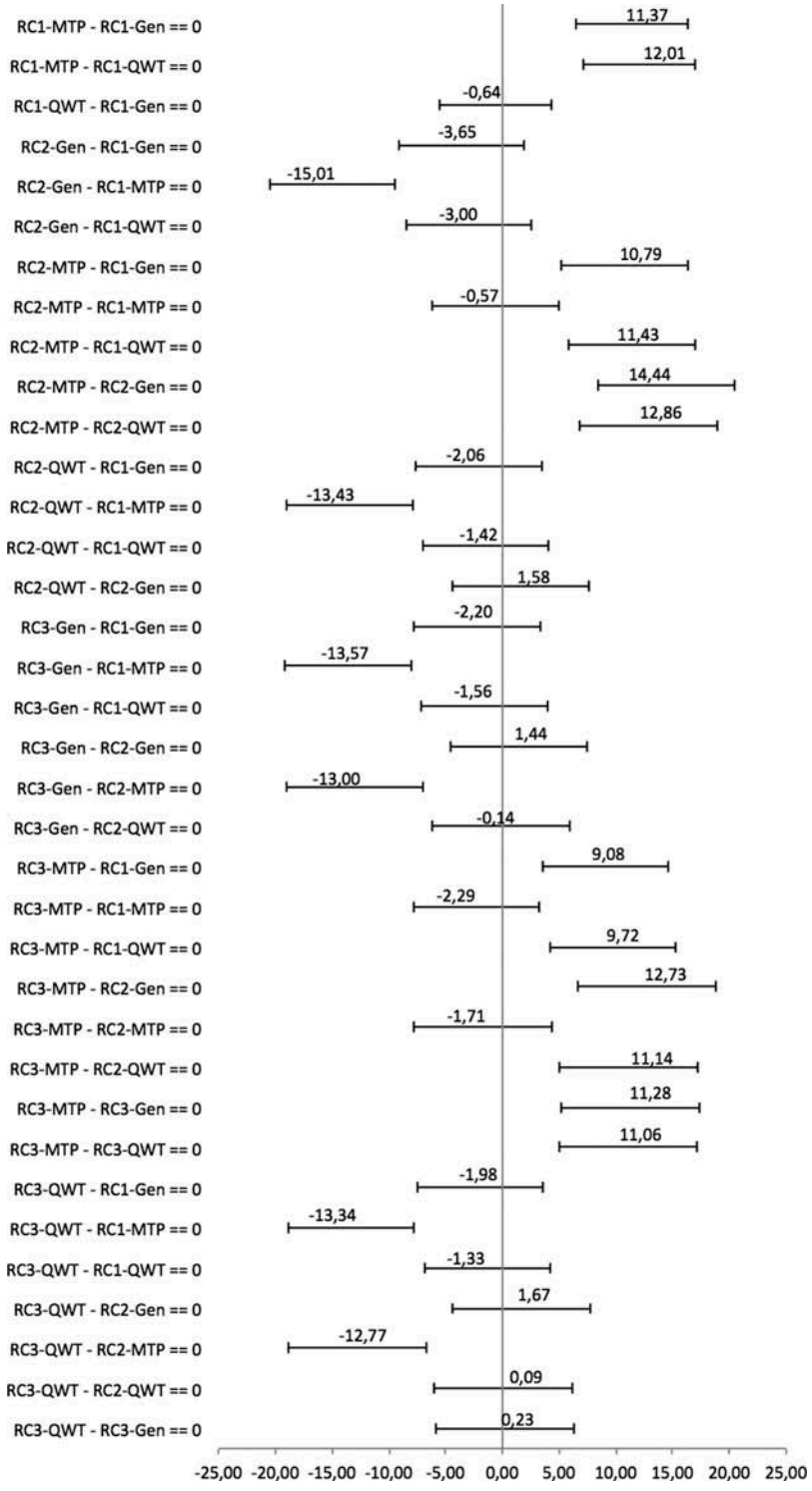

FIG. 19. Pairwise comparisons of error rates for each method and remote control (95% family-wise confidence level).

Although articles such as Orbist et al. (2008), recommend not to rely on these devices, and the results of some experiments such as those in MacKenzie et al. (2009) may not be as were expected, it is increasingly evident that the popularity of devices to interact with televisions different to the conventional remote controls will increase in the near future.

Thus, experiments may be undertaken to compare the performance of the evaluated methods with devices such as gyroscopic remote controls (pointing devices), mini-keyboards, touchpads, tablets, smartphones, or multitouch devices. Apart from writing speed and accuracy, we would also like to measure other factors that may influence the experience of users. By incorporating further equipment, we would like to measure stress levels, brain activity, eyesight effects, and similar parameters, which may allow us to draw even more interesting conclusions.

FUNDING

This work was partially supported by the University of Oviedo and the Principality of Asturies through the project SV-PA-13- ECOEMP-75 and by the cable operator Telecable de Asturias S.A.U. through the project "Diseño de sistema para gestión y provisión de aplicaciones/servicios de televisión digital interactive."

REFERENCES

Aoki, R., Maeda, A., Watanabe, T., Kobayashi, M., & Abe, M. (2010). Twist tap: Text entry for TV remotes using easy-to-learn wrist motion and key operation. *IEEE Transactions on Consumer Electronics*, *56*, 161–168.

Bellman, T., & MacKenzie, I.S. (1998). A probabilistic character layout strategy for mobile text entry. *Proceedings of Graphics Interface ' 98*, 168–176.

Brandtzæg, P. B., Heim, J., & Karahasanović, A. (2011). Understanding the new digital divide—A typology of Internet users in Europe. *International Journal of Human–Computer Studies*, *69*, 123–138.

Brewbaker, C. R. (2008). *Optimizing stylus keyboard layouts with a genetic algorithm: Customization and internationalization*. Department of Computer Science, Iowa State University, Ames.

Butts, L., & Cockburn, A. (2002). An evaluation of mobile phone text input methods. *Journal of Australian Computer Science Communications*, *24*, 55–59.

Choi, S., Han, J., Lee, G., Lee, N., & Lee, W. (2011). RemoteTouch: Touchscreen-like interaction in the TV viewing environment. *Proceedings of CHI 2011*, 393–402.

Clarkson, E., Clawson, J., Lyons, K., & Starner, T. (2005). An empirical study of typing rates on mini-QWERTY keyboards. *Proceedings of CHI '05*, 1288–1291.

Congressional Record, Vol. 155. (2009). DTV Delay Act, Public Law 111-4-Feb. 11, 2009.

Crawley, M. J. (2007). *The R book*. Chichester, England: Wiley & Sons.

European Broadcasting Union. (2006). Digital Video Broadcasting (DVB); Multimedia Home Platform (MHP) Specification 1.0.3. ETSI specification ETSI ES 201 812 V1.1.2.

European Telecommunications Standards Institute (ETSI) ES 202 130 V2.1.2, 2007-09, Human Factors (HF); User Interface; Character repertories, ordering rules and assignments to the 12-key telephone keypad. http://portal.etsi.org/stfs/STF_HomePages/STF300/es_202130v020102p.zip

European Union. (2005). *Communication from the Commission to the Council, the European Parliament, the European Economic and Social committee and the Committee of the Regions on accelerating the transition from analogue to digital broadcasting* {SEC(2005)661}.

Gargi, U., & Gossweiler, R. (2010). QuickSuggest: Character prediction on web appliances. *Proceedings of WWW 2010*, 1249–1252.

Geleijnse, G., Aliakseyeu, D., & Sarroukh, E. (2009). Comparing text entry methods for interactive television applications. *Proceedings of EuroITV'09*, 145–148.

Iatrino, A., & Modeo, S. (2006). Text editing in digital terrestrial television: A comparison of three interfaces. *Proceedings of EuroITV'06*, 224–241.

Ingmarsson, M., Dinka, D., & Zhai, S. (2004). TNT—A numeric keypad based text input method. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 639–646.

Költringer, T., Isokoski, P., & Grechenig, T. (2007). TwoStick: Writing with a game controller. *Proceedings of Graphics Interface 2007*, 103–110.

MacKenzie, I. S. (1991). *Fitts' law as a performance model in human-computer interaction* (Unpublished doctoral dissertation). University of Toronto, Toronto, Ontario, Canada.

MacKenzie, I. S., Kober, H., Smith, D., Jones, T., & Skepner, E. (2001). LetterWise: Prefix-based disambiguation for mobile text input. *Proceedings of the ACM Symposium on UIST*, 111–120.

MacKenzie, I. S., Lopez, M. H., & Castelluci, S. (2009). Text entry with the Apple iPhone and the Nintendo Wii. *Proceedings of CHI2009*.

Nesbat, S. B. (2003). A system for fast, full-text entry for small electronic devices. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 4–11.

Nielsen, J. (2012). *Remote control anarchy*. Available from http://www.useit.com/alertbox/20040607.html

Oniszczak, A., & MacKenzie, I. S. (2004). A comparison of two input methods for keypads on mobile devices. *Proceedings of the Third Nordic Conference on Human-Computer Interaction*, 101–104.

Orbist, M., Bernhaupt, R., & Tscheligi, M. (2008). Interactive TV for the home: An ethnographic study on users' requirements and experiences. *International Journal of Human-Computer Interaction*, *24*, 174–196.

Perrinet, J., Pañeda, X. G., Cabrero, S., Melendi, D., García, R., & García, V. (2011). Evaluation of virtual keyboards for interactive digital television applications. *International Journal of Human-Computer Interaction*, *27*, 703–728.

RAE - Royal Spanish Academy. (2001). *Spanish language dictionary* (22nd ed.). Available from http://www.rae.es/rae.html

Rick, J. (2010). Performance optimizations of virtual keyboards for stroke-based text entry on a touch-based tabletop. *Proceedings of the 23nd Annual ACM Symposium on UIST*, 77–86.

Ritter, F. E., & Schooler, L. J. (2002). The learning curve. In *International encyclopedia of the social and behavioral sciences* (pp. 8602–8605). Amsterdam, the Netherlands: Pergamon. Available from http://www.iesbs.com/

Siek, K. A., Rogers, Y., & Connelly, K. H. (2005). Fat finger worries: How older and younger users physically interact with PDAs. *Proceedings of Interact 2005*, 267–280.

Silfverberg, M., MacKenzie, I. S., & Korhonen, P. (2000). Predicting text entry speed on mobile phones. *Proceedings of the CHI2000 Conference*, 9–16.

Spira, J. B. (2011). Internet TV: Almost ready for prime time [Tools & Toys]. *IEEE Spectrum*, *48*, 24–26.

Sporka, A. J., Polacek, O., & Slavik, P. (2012). Comparison of two text entry methods on interactive TV. *Proceedings of the 10th European Conference on Interactive TV and Video*, 49–52.

Taveira, A. D., & Choi, S. D. (2009). Review study of computer input devices and older users. *International Journal of Human–Computer Interaction*, *25*, 455–474.

Varcholik, P. D., LaViola, J. J., Jr., & Hughes, C. E. (2012). Establishing a baseline for text entry for a multi-touch virtual keyboard. *International Journal of Human-Computer Studies*, *70*, 657–672.

Vega-Oliveros, D. A., Pedrosa, D. C., Pimentel, M. G. C., & De Mattos Fortes, R. P. (2010). An approach based on multiple text input modes for interactive digital TV applications. *Proceedings of the 28th ACM International Conference on Design of Communication*. 191–198.

Wigdor, D., & Balakrishnan, R. (2004). A comparison of consecutive and concurrent input text entry techniques for mobile phones. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 81–88.

Wobbrock, J. O., Myers, B. A., & Aung, H. H. (2004). Writing with a joystick: A comparison of date stamp, selection keyboard, and EdgeWrite. *Proceedings of Graphics Interface '04*, 1–8.

Zhai, S., Hunter, M., & Smith, B. A. (2000). The Metropolis keyboard—An exploration of quantitative techniques for virtual keyboard design. *Proceedings of ACM Symposium on User Interface Software and Technology*, 119–128.

ABOUT THE AUTHORS

Aurora Barrero is a Computer Science Engineer from the University of Oviedo and a Ph.D. student with an interest in the area of multimedia systems and services, content distribution networks, and interactive digital TV services. She is a Research Engineer in the Department of Computer Science of the University of Oviedo.

David Melendi is a Computer Science Engineer with a Ph.D. from the University of Oviedo and an interest in multimedia systems and services, content distribution networks, idTV services, and mobile ad hoc networks. He is an Associate Professor at the University of Oviedo and a member of the W3C.

Xabiel G. Pañeda is a Computer Science Engineer with a Ph.D. from the University of Oviedo and an interest in multimedia systems and services, content distribution networks, idTV services, and mobile ad hoc networks. He is an Associate Professor at the University of Oviedo and a member of the W3C.

Roberto García is a Telecommunications Engineer from The Technical University of Madrid with a Ph.D. from the University of Oviedo and an interest in telecommunication networks and services, applied to performance analysis, modelling, and simulation of systems and services. He is an Associate Professor at the University of Oviedo.

Sergio Cabrero is a Telecommunications Engineer from the University of Oviedo and a Ph.D. student with an interest in the area of telecommunication networks, interactive digital TV services, multimedia services, and mobile ad hoc networks. He is a Teaching Assistant in the Department of Computer Science of the University of Oviedo.