Speech Enhanced Remote Control for Media Terminal

Aseel Ibrahim*, Jonas Lundberg & Jenny Johansson*

*Nokia, Diskettgatan 11, SE-583 35 Linköping, Sweden
Department of Computer and Information Science, SE-581 83 Linköpings University, Sweden
{aseel.ibrahim, jenny.johansson}@nokia.com, jonlu@ida.liu.se

Abstract

A media terminal box combines digital television and services on the World Wide Web. This device will be available in many homes and the interaction with it occurs via a remote control and a visual presentation. The problem is the navigation difficulties among the huge number of television channels. The aim of this study is to investigate whether spoken commands could solve the navigation problem. In this study two input techniques were tested: remote control and speech input. The results showed that speech input was more effective as steps to complete tasks were less and shortcuts were used more often in the speech condition. However, the subjective data showed that the subjects were more satisfied with the remote control input. In conclusion, we recommend multimodal interaction where of speech input to complement the remote control unit.

1. Introduction

The new digital media channels, including web services, television, and games, will be delivered to the homes of ordinary people. When delivered, the users must be able to find and select the desired channel. A media terminal box, which can be connected to devices like the television set, has been constructed to solve this problem, allowing the user to navigate between the channels using a traditional remote control (RC) and a visual presentation of the available channels called Navibars. The Navibars consists of a cross with two bars, one vertical bar and one horizontal bar (see Figure 1). The horizontal bar displays categories, one category in each box. Each box contains a number of containers. The containers gives access to various services, television channels, television programs, Internet content, games, e-mail applications, music, etc. When a box is in the focus the content of it is displayed in the vertical bar.

The problem with this solution is that navigation in the huge space of channels is difficult due to several reasons. First, the names and positions of all channels can’t be presented on-screen at the same time, which makes it hard to find the desired channel. Second, there can be a long distance between the current channel and the desired channel. These are the problems that will be addressed in this paper.

Using speech to name the channels would let the user reach the desired channel directly, using the name as a spoken shortcut. In this paper we compare shortcuts and navigation by means of traditional RC with shortcuts and navigation by means of spoken commands. As shortcuts, using a traditional remote control, are restricted to numbers, we predicted that spoken commands would be an advantage, as the novice user would be able to guess the names of channels, and that experienced users would have an easier time remembering the names of channels.

2. Method

The aim of the study is to investigate whether speech input is suitable in the media terminal domain, whether the media terminal can be controlled by means of speech input, and whether the user would be satisfied if speech interaction is provided.

Considering the vast number of television channels a media terminal can have, the subjects will probably find it hard to navigate by means of the RC. We believe that speech input could provide more efficient interaction. Efficiency is defined here as completing a specific task with as few steps as possible. Navigating through Navibars to move to a desired channel could take many steps. Therefore we believe that spoken commands are likely to result in substantial efficiency...
gains as spoken commands allow the user to use shortcuts, a quicker way to move to the desired channel by uttering the name of the channel. Shortcuts are also provided in the RC condition by typing the number of the channel, but they require more mental effort as the user has to learn and remember them. Therefore we believe that the users will be more comfortable and express more satisfaction with using the spoken commands.

In our study we defined the following hypothesis: users prefer shortcuts to navigate and speech input will allow the use of shortcuts in situations where RC input users will have to resort to navigation. Speech input users will therefore use shortcuts more frequently than RC users, and be more satisfied with the system.

2.1. Subjects
Sixteen subjects participated in the experiment and they were chosen to fit to the target group. Eight of them were male and eight were female. The age of the subjects ranged from 17 to 60 years with an average of 38 years.

A between groups repeated measures factorial design was used in the experiment. The subjects were divided into two groups: a RC group that tested the RC condition, and a speech command group that tested the speech command condition. Eight subjects participated in each group.

2.2. Prototype
A Navibars prototype with two conditions was developed for the experiment. The interaction in the first condition occurred by means of key RC and with the second condition by means of speech input. The Navibars prototype consisted of nine boxes with a maximum of nine containers in each box. To adjust the prototype for the study, shortcuts were built into Navibars. The shortcuts were only provided for the containers in the vertical bar. Each box in the horizontal bar was numbered, for example the first box, "Family TV" was numbered 1 and the third box "WEBB" was numbered 3 (see Figure 1). The containers in each box were numbered as well. In order to know which box each container belonged to, the box number was positioned to the left of the container number. For example, the first container in the "WEBB" box was numbered 31 and the second 32.

The subjects in the RC condition interacted with Navibars by using arrow keys, a select button, a Navibars button, and 0 to 9 digit buttons. The subjects accessed Navibars by pressing the "Navibars" button. The navigation was controlled by pressing the up, down, left, and right arrow keys. When a specific container was in focus, it could be chosen by the select button. The shortcuts were used by entering the number for the desired container. All subjects' actions were logged. The logs included information about the type of button pressed by the subject.

A Wizard of Oz was used as a speech recognizer and simulated a keyword matching system. The interaction occurred by uttering voice commands. A very simple speech RC was provided and it consisted of only one button, the speech button. The subject had to press this button in order to activate the speech recognizer, in this case the Wizard. The commands for navigation were up, down, right, and left. The shortcuts were obtained by uttering the number to the particular container. The subject could also use the channel logotype as voice command. The voice command for selecting a specific content was "select" and for bringing up Navibars, was "Navibars". All Wizards' actions, which were equivalent to the subject's actions were logged. The logs included information about the type of the button pressed by the Wizard.

2.3. Procedure
The experiment consisted of five parts; a pre-test questionnaire, a training session, the test of the prototype, a post-test questionnaire, and finally a post-test interview.

First, the subjects were asked to fill out a pre-test questionnaire. The purpose of the pre-test questionnaire was to collect demographic information about the subjects. Next, each subject went through a training session. The subject was given some scenarios that included tasks for navigation and for using shortcuts.

After the training session, the test started. There were fourteen task scenarios that were presented in the same order for all subjects. Each task was presented to the subject by the experimenter who read it aloud. The experimenter observed the subject during the test and made notes of whether the subject used shortcuts or navigated in order to complete the task.

After the test, the subject was asked to fill out a post-test questionnaire where the subject's satisfaction could be ranged on a scale from one to five, from strongly agree to strongly disagree. Finally, a post-test interview was undertaken with each subject. The interview was intended to yield more qualitative assessments of system learnability as well as satisfaction, experience from talking to the system, and willingness to use the system.

2.4. Task Scenarios
Fourteen task scenarios were used in the study. The aim of the tasks was to find television/Internet content, or services by navigating through Navibars. The scenarios corresponded to various situations from real life, where both novice and experienced users were simulated. The scenarios in Table 1 are translated from Swedish.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A documentary with title Viability will run on TV during the morning. Look up channel TV2 by using Navibars and watch the program.</td>
</tr>
<tr>
<td>S2</td>
<td>You want to check if you have received any new e-mail. Go to <a href="http://www.hotmail.com">www.hotmail.com</a>, which is available in the box My Things. Use the Navibars.</td>
</tr>
<tr>
<td>S3</td>
<td>Kajsa is back home and she wants to watch a children program on channel Cartoon Network that is available in box Children TV. Use Navibars to go there!</td>
</tr>
<tr>
<td>S4</td>
<td>Kajsa has homework about penguins and needs to collect information. She could use the web site <a href="http://www.altavista.com">www.altavista.com</a>. Go to this web site with help of Navibars!</td>
</tr>
<tr>
<td>S5</td>
<td>Move to TV channel TV6 by using the number of the channel. Use Navibars to find out the number if you do not know it already.</td>
</tr>
<tr>
<td>S6</td>
<td>Stefan thinks that a reprise about trains on TV channel SVT1 is running right now. SVT1 is available on number 11, go there by using the channel number!</td>
</tr>
<tr>
<td>S7</td>
<td>It's dinnertime and you want to hear some background music while you are eating. Go to MTV that has number 52 by using the channel number!</td>
</tr>
<tr>
<td>S8</td>
<td>MTV was boring maybe you can find something better on ZTV. Go there, as you prefer!</td>
</tr>
<tr>
<td>S9</td>
<td>Stefan has just came home and he wants to see sport on TV channel Eurosport that is available on channel number 92. Go there, as quickly as possible as you prefer!</td>
</tr>
<tr>
<td>S10</td>
<td>Stefan now wants to go to the cinema and he wants to check which movies are available. There is a movie database on the web site <a href="http://www.IMDb.com">www.IMDb.com</a> that is available on number 73. Go there, as you prefer!</td>
</tr>
<tr>
<td>S11</td>
<td>Today is the tenth of March and a movie with title...</td>
</tr>
</tbody>
</table>
The purpose of the first four scenarios was to introduce the use of Navibars. The subject was requested to complete the tasks by using the Navibars. Arrow keys were used in the RC condition and voice commands (up, down, right, and left) in the speech condition.

In both scenarios 6 and 7, the subjects were asked to move to a specific channel by pressing/uttering the number of it. The number was given to the subjects. With these scenarios shortcuts were introduced.

In the last seven scenarios, the subjects were asked to complete the tasks in their preferred way. The purpose of scenarios 8 and 9 was to simulate a subject knowing both the name of the channel/web site as well as the number of it in order to investigate which way the subject would prefer to perform the task. In the following two scenarios, the subjects were given a daily newspaper that was used to perform the task. The context of these scenarios was that the subjects were a novice and had a daily newspaper, they noticed that an interesting movie was going on, and wanted to go to the desired channel. The purpose of the scenario was to find out how the subjects would act when they knew the name of the channel but they did not know where it could be found.

The purpose with scenario 14 was to investigate whether shortcuts would be useful to provide for the boxes. In the last scenario the name given to the subjects were not the same as the channel logotype simulating a user having a different name for a channel than the name of understood by the system.

3. Results

3.1. Steps to Complete Tasks

For each subject, the number of steps to complete the tasks was measured. The way subjects conducted the tasks could also be extracted from the logs.

The results showed that when navigating in the system the subjects in the speech condition overall used significantly fewer steps than the RC condition. The total number of steps used in the RC input was 951 while the total number of steps in the other condition was 623. This difference was statistically significant (t(14) = 0.0065, p < 0.05, one-tailed).

The result of scenario 3 showed the difference in number steps needed to perform a task when the subject chose a "wrong" or long way to the target. The number of steps in the RC condition was 51% more. This difference was statistically significant (t(14) = 0.0005, p < 0.05).

3.2. Usage of Shortcuts

The frequency of using shortcuts could be measured in all scenarios. This information was extracted from the test logs.

The results showed that shortcuts were more frequently used in the speech condition. The total number of shortcuts in the speech condition was 55 while the total number in the RC condition was 32. This difference was statistically significant (chi-square(1) = 0.0015, p < 0.05).

In scenario 8, the subjects in the speech command group used shortcuts more frequently. The difference was statistically significant (t(8) = 0.0025, p < 0.005, one-tailed). Only one subject in the speech condition used the up, down, right, and left keys to navigate while the others used shortcuts. In the other condition, all subjects used the arrow keys, see Figure 2. The difference was statistically significant (chi-square(1) = 0, p < 0.05).

Shortcuts were also more often used in both scenarios 11 and 12. The results showed that there was a significant difference in scenario 11 (t(11) = 0.001, p < 0.05, one-tailed), but not in scenario 12 (t(7) = 0.6, p < 0.05, one-tailed). Scenario 12 in Figure 2 shows that seven subjects used shortcuts in the speech condition while no subject in the other condition. This difference was statistically significant (chi-square(1) = 0, p < 0.05).

3.3. Subjective User Evaluation

The results from the subjective post-test questionnaire as well as from the semi-structured, post-test interviews were analyzed. The subjects questionnaire answers ranged from one (corresponding strongly agree) to five (corresponding strongly disagree).

The results from the questionnaire showed that overall, subjects expressed satisfaction with the prototype of the media terminal. However, this was somewhat more obvious in the RC condition. The mean satisfaction among subjects in the speech command group was 1.953, while in the other group it was 1.5. This difference was statistically significant (t = 0.0125, p < 0.05, one-tailed).

All subjects agreed that the shortcuts were easy to use in RC condition. The mean agreement was 1. From the semi-structured interviews the subjects remarked that the shortcuts would become even more natural to apply once different numbers had been learned. However, one subject remarked that when the number of channels increased, it might be hard to remember the respective numbers.

Subjects in both groups thought that it was easy to find items in the system. The mean for the RC condition was 1 while the other condition was 2. According to a one-tailed t-test, this difference was not significant (t = 0.111, p > 0.05).
Subjects in the speech command group felt that they could control navigation by means of voice (mean was 3). Alternatively, several subjects remarked at the post-test interview that they felt a bit unaccustomed talking to the media terminal. One subject commented that it would be hard to use the voice control when watching television together with others, since they thought that the voice interaction demands silence. When subjects were asked if they wanted a RC unit rather than speech input, five of them agreed, one responded negatively, and the last subject wanted both. Moreover, even if most subjects would like to have the media terminal at home, some of them meant that it would be useful in the future, but that technology has to advance before the voice control will become practically feasible.

4. Discussion

The results verified our hypothesis. Shortcuts were more often used in the speech condition, which resulted in fewer steps required to perform a task. The distance between the current position and the desired position affected the gain of using shortcuts. If the distance was not large enough, the gain was also small.

This usage of shortcuts did contribute to the overall effect of fewer steps taken in the speech condition. However, a random effect did contribute as well. In task scenario 3, more subjects in the RC condition took the wrong way, while navigating to the target. The target was next to the starting point, to the left. In the RC condition most subjects instead went to the right, which resulted in eight steps to reach the target goal, instead of one. Rather than being an effect of speech or RC, this is a problem facing every system of this kind, and the easiest way to avoid the effect is not to use too many boxes. As the system probably will have too many boxes, we suggest the use of shortcuts, which will solve the problem as well, and which is the next topic of this discussion.

In task scenario 11 and 12, the subject was given a television listing for some channels, and the name of a movie to see. As we had predicted, the speech command group managed to guess the name of the channel, from the channel logotype in the television listing, and use it as a shortcut. The RC group, however, were unable to guess the number of the channel and thus resorted to navigation to reach it. This was significantly less efficient for task scenario 11, but not for task scenario 12, since in task 11 the distance between the current and the desired channel when navigating was eight steps, compared to three steps in task 11. The same result was achieved in task 8, simulating a user knowing the name of the desired channel, but not its number. Again the speech command group used the name as a shortcut while the RC group had to resort to navigation.

Scenario 14 illustrated that the names of channels must be what the user expects them to be. Here, the subject was given a name of a channel, different from the name understood by the system. The speech command subjects tried to use that name as a shortcut, but they did not try a different one when that failed. This means that the common names for each channel including nicknames should be included in the speech recognizer’s lexicon.

The experiment thus shows that users prefer shortcuts to navigation, and that speech input allows more intuitive shortcuts than the numbers allowed by a traditional RC. We do consider speech input a more elegant solution than to replace the RC unit with a complete keyboard. With regards to the scale of the prototype system, which could have a maximum of 9 boxes with 9 channels in each box, we believe that the advantages of speech shortcuts will be even greater, with a full scale system allowing a hundred boxes or more.

Having established a preference for speech input, from the quantitative analysis, the post-test questionnaire and interviews indicate that the RC is still necessary. The subjects of both input types were satisfied with both the media terminal, the ease of finding channels and the use of shortcuts. The subjects in the speech condition on the one hand felt that they could control the media terminal by voice, but on the other hand felt slightly uncomfortable doing so, and expressed a wish to have a traditional RC in addition.

In conclusion, the study showed that subjects preferred to use shortcuts rather than to navigate, and that the speech command group could do so in situations where the RC group had to resort to navigation. Therefore we recommend speech control for this kind of system. As the questionnaire and interview showed users preferred to have a RC unit, we recommend speech control as a complement to the traditional RC unit, rather than as a substitute for it. This is based on the strategy of developing multimodal user interfaces where the strengths of one mode is used to overcome the weaknesses in another mode.

5. References


