Voice Command Generation for Teleoperated Robot Systems

M. Ferre
FAIS - EUITI
Universidad Politécnica de Madrid
Madrid, 28012

J. Macías-Guarasa
DIE - ETSIT
Universidad Politécnica de Madrid
Madrid, 28040

R. Aracil
DISAM- ETSII
Universidad Politécnica de Madrid
Madrid, 28006

A. Barrientos
DISAM- ETSII
Universidad Politécnica de Madrid
Madrid, 28006

Abstract

In a teleoperated system, several degrees of freedom (DOF) are controlled by the operator, in which different control levels occur to carry out remote tasks. These control levels are implemented by using two kinds of commands: low level commands (LLCs), and high level commands (HLCs) [1].

LLCs are related to tasks carried out on joint robots. In this paper, an experiment comparing voice input to traditional input devices, master-arm and joystick, has been conducted.

As for HLCs, human-robot interaction by means of voice has also been implemented by using two types of interfaces. The first one presented under a menu form: a set of commands is displayed for the operator to choose among them. The second one is based on natural language processing techniques, so that voice commands are generated following a given imperative grammar structure.

The speech recognition engine is based on an isolated word recognition system [2], designed to work in real time with a minimum hardware add-on. It converts the spoken utterances into text, as the first step to commands generation to control the remote devices.

It is shown how voice input is an efficient technique for HLCs, whereas traditional input devices performs are more appropriate for LLCs.

Some of the main results of this work have already been successfully applied to a teleoperated system devoted to power live line maintenance, called ROBTET [3].

1 Introduction

Teleoperated systems are controlled by human operators, who after having received the appropriate information continuously generate commands in a teleoperated task.

Verbal communication is the primary and most natural way of human communication, so that voice commands are an eminently suitable technique for commanding teleoperated devices.

Different approaches to design a voice interface are available for the system engineer. Careful attention should be paid to evaluate them, as compared to traditional ways of interaction in the task under research and development.

In general, voice interfaces are based on a translation system from operator utterance sentences into teleoperated device commands. An efficient voice recognition system would lead to an improvement in teleoperation task performance. The question is to decide what “efficient” means in the context of the application we are pursuing. As it is later stated, the main requirement is a simple and straightforward voice-command relationship.

2 The speech recognition engine

The speech recognition process is performed in a low cost dedicated DSP board [4], including voice acquisition, speech preprocessing and the recognition process itself. A carefully designed speech API has also been implemented to facilitate the development and test of different approaches.

The engine, originally designed as a preselection stage [5], achieves speaker dependent isolated word recognition for medium to large vocabulary sizes. It follows a bottom-up, two-stage strategy, as shown in figure 1 [2]. Its main components are a feature extractor, a phonetic string build-up module (PSBU), and a lexical access stage (LA). The PSBU decides the optimum concatenation of the allophones in the utterance, and the LA computes similarities between the phonetic string and each entry of the dictionary [6], offering a final list of candidate words.

The static modeling [5] used in the PSBU stage exploits the phonetic-acoustic characteristics of Spanish language, which can be reasonably modeled considering only stable portions of speech.

The search in the LA is implemented as a tree to save computational effort, and the algorithm is based on a dynamic programming procedure, where substitution, insertion and deletion alignment errors are considered. Taking all these facts into account, the computational demands are really low, so that extension to larger vocabularies is not a problem.

The API allows easy implementation and test of

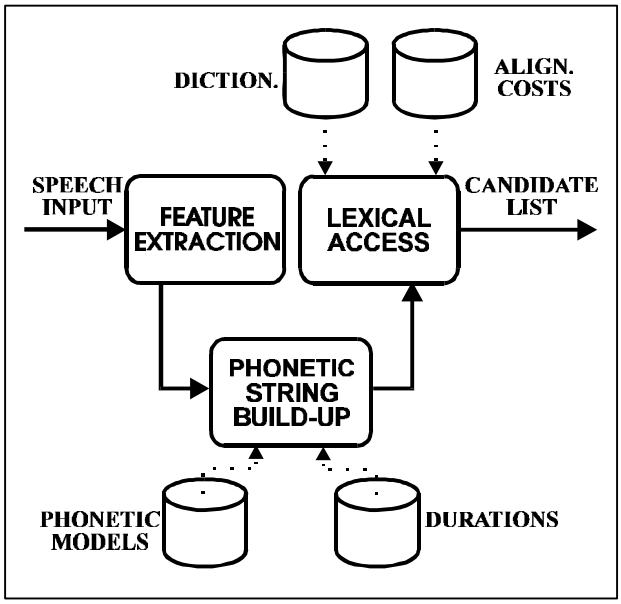

Figure 1: Speech Recognition Engine Architecture

different vocabularies. Restriction methods are also included, in order to improve accuracy, mainly based on sub-vocabulary selection.

Additional tools have been developed, mainly to allow different speakers to train the system and generate the speaker dependent acoustic models needed for better recognition accuracy. The enrollment procedure consists of reading only 150 words, and takes about 20 minutes.

3 Experiments to compare operator input devices generating LLCs

3.1 Problem description

Guidance is the lowest control level in telerobotics. It provides the trajectory position and reference marks to be followed by the robot, i.e. grips or tools, as the operator guides teleoperated devices on remote task execution.

In these experiments, a PUMA robot is teleoperated via the Alter communication line, and different operator input devices have been used for the guiding process.

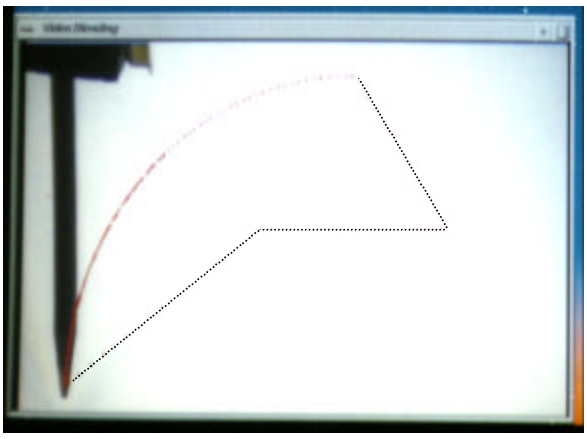

Figure 2. LLC generation experiment

Several research works have been developed to compare rate control and position control, since the operator moves the master, and master join positions are used to generate position references for slave robot. Some studies at the Jet Propulsion Laboratory [7], [8], concluded that position control provides a more efficient performance than rate control.

At the beginning of the 80s McRuer [9] proposed an operator behavior model for a guidance task using master-arms. According to McRuer, “operator plus process” can be modeled as an integrator with a delay. This is a well-accepted model, and for our

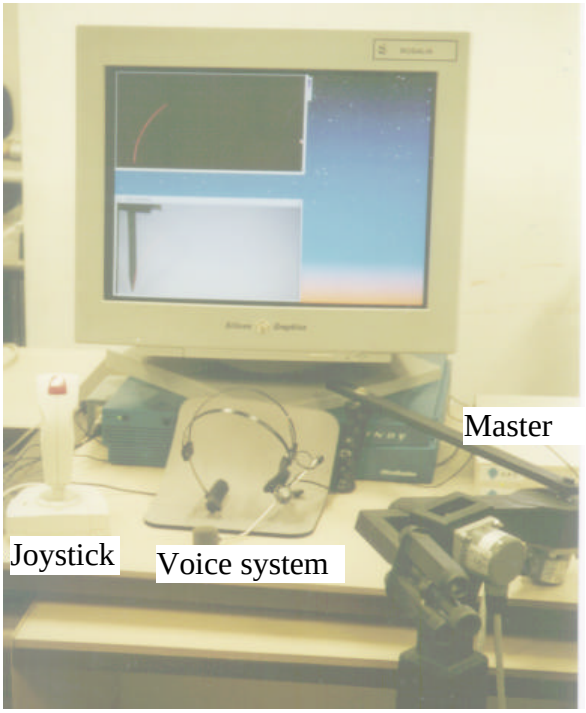

Figure 3: Operator input devices

study, it provides an adequate way to measure the guidance quality with different input devices. This measure is based on the distance between the ideal trajectory and the actually executed robot trajectory.

In the guidance experiment, a video image with the robot end position is displayed, and a point is continuously drawn over it, to show the operator the ideal robot trajectory. The distance between the robot end and the last point drawn is the actual error for every trajectory step. Trajectory quality is measured as the average error among all trajectory steps.

Figure 3 shows an example of the defined trajectory for this experiment with the robot end placed at the last point. The trajectory has only 2 DOF, since it is defined on a plane and the video image is parallel to this trajectory plane.

3.2 Operator input devices

Master-arm, joystick, and word recognition systems have been compared in this experiment. All the peripherals used are shown in figure 3.

The devices have been evaluated with a similar level of proprioception. Devices with an inferior proprioception are penalized [10].

The master-arm used, designed at the DISAM Dept., has 3 DOF, with an anthropomorphic configuration. Only 2 DOF were used in this experiment to generate position references, both of them confined within a horizontal plane.

The joystick has 2 DOF and it generates rate references for the teleoperated robot. Like the master-arm, joystick movements are congruent with the feedback video image, so that left joystick movements generate left robot velocity, and so on.

Figures 4 and 5 show operators executing the LLC experiments using master-arm and joystick.

The word recognition system is implemented by the VISHA system. The operator instructs tasks following given words, as shown in figure 6. In this

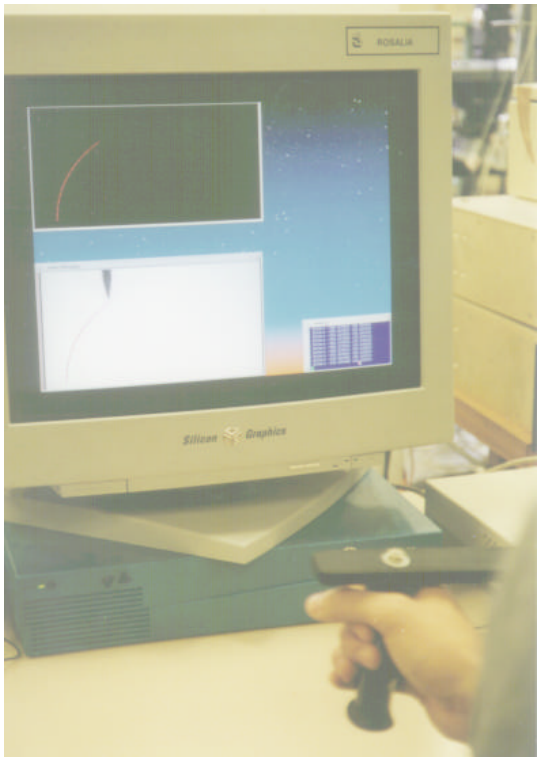

Figure 4: Master-arm execution

experiment, a dictionary composed of 13 words has been designed, which corresponds to numbers from “one” to “twelve”, clockwise arranged, and “ stop”. Numbers indicate movement direction, so that, for example, “twelve” means up, and “six” down. Other strategies may be applied to this voice experiment, as long as it is a straightforward designed system, that is, the relationship between commands and verbal utterances is simple enough not to make the operator task more difficult. Considering the limited

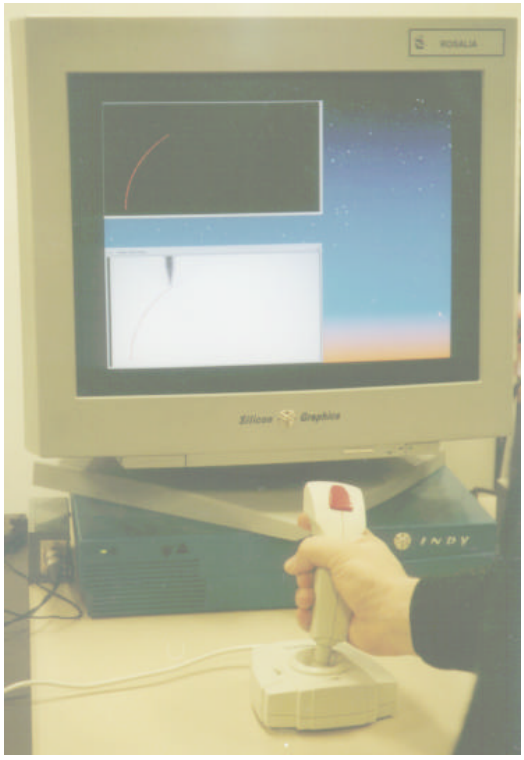

Figure 5: Joystick execution

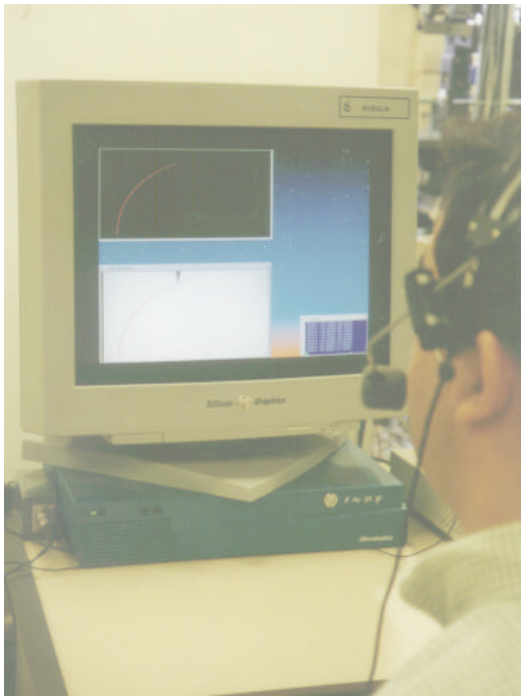

Figure 6: Voice execution

vocabulary in this task, a 98% recognition rate was achieved, high enough to ensure the recognition engine errors do not interfere with the results and conclusions discussed below.

In this experiment, the trajectory is implemented with different speeds. In the case of voice input speed was chosen to be equal to the trajectory maximum speed. Then, on the trajectory execution, it is necessary to stop the robot in order to describe the given trajectory with the proper speed.

Alternative mechanisms could have been designed and implemented, using other strategies, but it is always necessary to design a straight one, so that its influence on the results is minimal. Otherwise, the voice system will not work properly, as the operator would have to translate actions into prescribed words, leading to performance degradation. For example, speed could even be indicated by voice, but in this case, the voice experiment would became more complex, and performance would decrease again, since two words would be needed to move the robot,

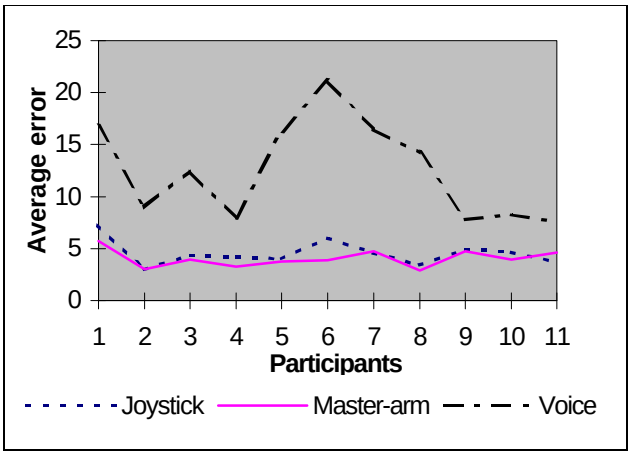

Figure 7: Graphical results for LLC

3.3 LLCs experiment results

Table 1 and figure 7 show the experiment results. Columns indicate the average error for each device. According to McRuer's research about master-arms. The average error measures the degree of operator and task achievement using different masters.

Analysis of the variance allows studying the influence of several factors in experiments. The key factors in the execution of the designed experiment are the input device and the operator. Influence of these factors can be studied by means of the F of

Fisher test. It compares if different samples come, or not, from the same population

Joystick	*Master-arm*	*Voice*
7,29	5,7	16,73
2,96	3,08	8,99
4,36	3,94	12,53
4,16	3,29	7,96
4,02	3,79	16,29
6	3,89	21,15
4,59	4,76	16,43
3,44	2,88	14,19
4,9	4,75	7,8
4,69	3,91	8,24
3,72	4,6	7,61
4,6	***4,1***	***12,5***

Table 1. LLC experiment results.

First, operators are compared. The Fisher test results show that there is not a significant difference between the operators *(F(10,20) = 1.65 < Critical Value (2.34))*. Therefore, as a conclusion, all operators have a similar behaviour for different input devices. Similarity between operators allows using the average value for each input device, because data come from the same population according the test.

The second analysed factor is the input device. The test shows significant differences between them (*F(2,30)=30.62 >> Critical Value (3.31)*). Test indicates the great difference in the execution time between devices, as shown in figure 7. These results make evident that performance varies in input devices, irrespective of operators.

A final test has been carried out between joystick and master-arm. Comparison between them indicates a similar performance (*F(1,21)=1.29 < Critical Value(3.31)*). Master-arm is connected with position control and joystick with rate control. Similarity between them proves that hand movements are more efficient than voice commands for LLC generation.

4 Voice HLCs generation

On HLC generation, two types of interfaces are studied: One based on a menu of options and the other in natural language processing techniques.

Ten operators, who subsequently were asked to draw conclusions, have tested both interfaces.

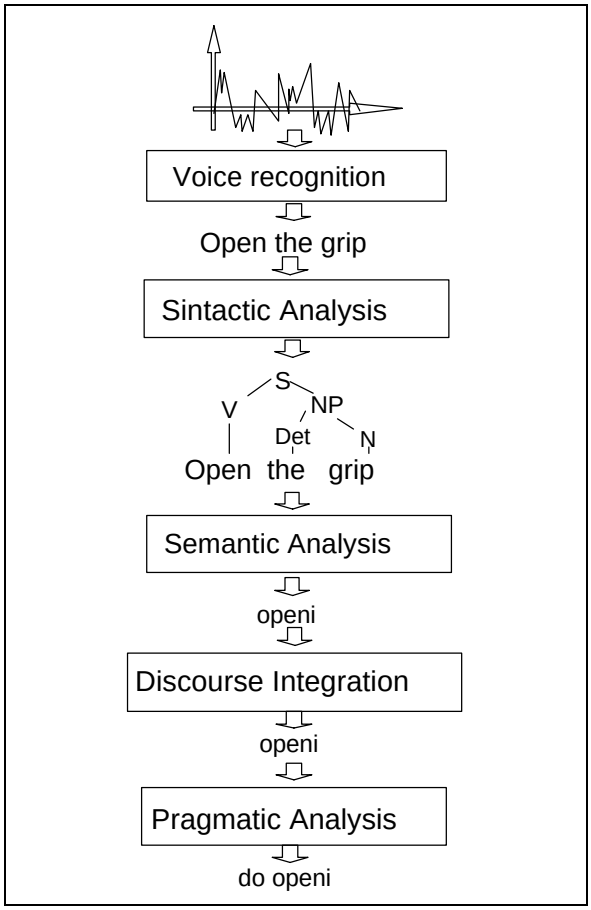

Figure 8: Natural language processing stages.

4.1 Natural language processing interface

Natural language is processed through different stages [11] as shown in figure 8. The first stage is again performed by the VISHA system, obtaining a list of candidate words. Next stage consists of syntactic analysis, for which an imperative grammar structure has been defined. At this stage recognition rate is improved from 82% to 96% due to its grammar category acceptance system. The following step is semantic analysis, in which the meaning is conveyed. Finally, discourse integration and pragmatic analyses are carried out to obtain the command expressed in the “target language”, in this case, instructions to a PUMA teleoperated robot (Val II).

4.2 Menu interface

A menu interface provides the operator with different options, available depending on the situation (state of the system), as shown in figure 9. The menu implementation is simpler than the one based on natural language processing, but, in any case, recognition accuracy should be quite similar as not to degrade relative performance of both approaches. A net of relationships has therefore been developed based on word linking, so that it reduces the menu alternatives offered to the operator at a certain time. A word will be accepted as long as the previous one has already been accepted.

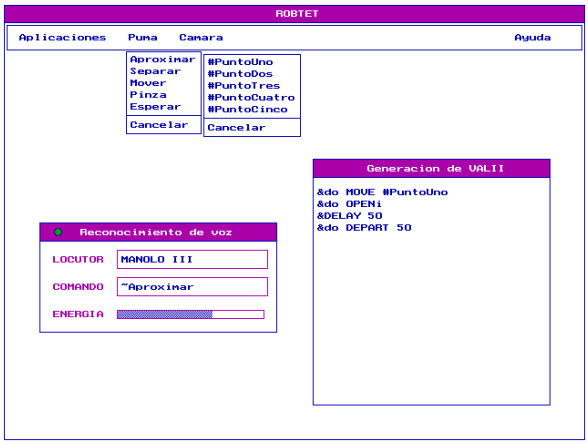

Figure 9: Menu implementation

4.3 Results

At the beginning of the experiment, the natural language processing interface was considered an attractive method and seemed to be the best approach. As the experiment was carried out, it turned out to be inadequate for the operator to execute tasks, since each word articulation has to be isolated, and the loss in fluency was quite considerable.

The menu interface was, however, more intuitive and easier for the operator since all the words to be used along the task are displayed on the screen and the interaction is based on a word-by-word scheme, which suits perfectly the isolated word recognition engine approach.

As for the natural language processing interface, the operator does not work comfortably if its recognition rate is under 90%. In a less noisy environment the VISHA system usually delivers a better performance, and recognition accuracy is increased using the restrictions described above. The menu interface further improves this performance resulting in an over 96% recognition rate.

5 Additional considerations on the speech interface

In real-world systems, non-technical, non-trivial question must be addressed, especially when interacting with the “human factor”, that may or not be technically proficient.

In this case, the recognition accuracy achieved has proved to be a key factor in the adequacy and acceptability of the voice interfaces designed.

The largest source of recognition errors, in a first approach, was due to the user’s “inexperience” when facing/dealing with an automatic speech recognition system. Great care had to be taken to teach users how to properly use the system and, even more important, how to proceed with the initial enrolment procedure, from which the acoustic models are extracted.

Anyway, the recognition performance obtained, in average, is not as high as expected, possibly due to the user’s lack of adaptation to the system, and this has been determinant for the results obtained. Careful examination of the interaction between these factors is mandatory.

6 Conclusions

After having analyzed the above-mentioned results, it can be concluded that a voice command system to control teleoperated devices is an efficient technique for HLC generation, tasks in which a single verbal command can be translated into a complex action involving several tasks.

The voice interface design must be adapted to the technological limitations of the recognition engine. Otherwise, the mismatch between the approach and the engine will make the technique inadequate, even though it may be, by itself, adequate.

The menu interface has proved to be an easy-to-apply approach, less complex than using natural language processing techniques. The former has been preferred by the test operators, mainly because

of the limitations of the recognition engine, based in isolated word recognition. In any case, a recognition rate over 95% is required for this voice interface to be accepted.

As for LLCs, hand movements (related to traditional input devices: master-arm and joystick) are more appropriate than speech control. The idea behind this is that the delay between the visual feedback the operators receive and the action this information produces is longer in the voice-input approach. It takes more time to utter a word than to move a hand, and the mental actions to actually utter the word are more complex than the ones required to move a hand. Then, hand movements are the natural way to guide a robot, and operator proprioception is the more important factor to achieve a good performance.

Considerations on users training time and expertise have not been discussed and should be addressed in the future.

In the speech-input modules, we are currently studying the possibility of including a state-of-the-art continuous speech recognition engine in the system. We firmly believe that the natural language processing techniques will show their power when applied in this case.

Some of the main results from this study have already been successfully implemented in a real world teleoperated system working on live power line maintenance, called ROBTET [3], in which the voice interface has been applied to control the cameras operation.

7 References

[1] Sheridan Tomas B, *Telerobotics*, Automation, and Human Supervisory Control. MIT Press. 1992.

[2] J. Macías-Guarasa, M.A. Leandro, J. Colás, Villegas, S. Aguilera and J.M. Pardo. *"On the Development of a Dictation Machine for Spanish: DIVO"*. Proceedings of the International Conference on Spoken Language Processing, pp. 1343-1346. Yokohama (Japan). 1994.

[3] R. Aracil, L. F. Peñín, M. Ferre, and A. Barrientos. *"ROBTET: A New Teleoperated System for Live-Line Maintenance"*. Proceedings of the 6th IEEE International Conference on Transmission and Distribution Construction, Operation and Live-Line Maintenance, ESMO-96. Columbus, Ohio, 1996.

[4] S. Aguilera, M.A. Berrojo, F.M. Giménez de los Galanes, J. Colás, J. Macías-Guarasa, J.M. Montero. *"Impaired Persons Facilities Based on a Multi-Modality Speech Processing System"*. Proceedings of the International Conference on Speech and Language Technology for Disabled Persons, pp. 129-132. Stocholm, May-June 1993.

[5] M. Leandro and J.M. Pardo. "Low cost speaker dependent isolated word speech preselection system using static phoneme pattern recognition". Eurospeech 1993, vol. 1, pp. 117-120. 1993

[6] L. Fissore et al. *"Lexical access to large vocabularies for speech recognition"*. IEEE Transactions on ASSP, vol. 37, n. 8, pp. 1197-1213. 1989.

[7] Kim Won S., Tendick Franck, Ellis Stephen and Stark Lawrence. 1987. *"A Comprison of Position and Rate Control for Telemanipulator Systems Dynamics"*. IEEE Journal of Robotics and Automation. Vol. 3, Nº 5.

[8] O´Hara John M. and Olsen Roy E. 1988. "*Control Device Effects on Telerobotics Manipulation Operations"*. Robotics and Autonomous Systems. Vol. 4. Nº 4.

[9] McRuer D. "*Human Dynamics in Man-Machine Systems"*. Automatica. Vol. 16. 1980.

[10] M. Ferre. Diseño de interfaces avanzadas para robots teleoperados. Desarrollo de un entorno de teleoperación con características multimedia. PhD thesis (In Spanish). Universidad Politécnica de Madrid (Spain). 1997.

[11] Crangle Collen, y Suppes Patrick. *Language and learning for robots*. Center for the Study of Language and Information. 1994.