A speech interface for Air Traffic Control terminals

J. Ferreiros, J. M. Pardo, R. de Córdoba, J. Macias-Guarasa, J. M. Montero, F. Fernández, V. Sama, L. F. d´Haro, G. González

***Abstract*— Several issues concerning the current use of speech interfaces are discussed and the design and development of a speech interface that enables air traffic controllers to command and control their terminals by voice is presented. A special emphasis is made in the comparison between laboratory experiments and field experiments in which a set of ergonomics-related effects are detected that can not be observed in the controlled laboratory experiments.**

The paper presents both objective and subjective performance obtained in field evaluation of the system with student controllers at an air traffic control (ATC) training facility. The system exhibits high word recognition test rates (0.4% error in Spanish and 1.5% in English) and low command error (6% error in Spanish and 10.6% error in English in the field tests). Subjective impression has also been positive, encouraging future development and integration phases in the Spanish ATC terminals designed by Aeropuertos Españoles y Navegación Aérea (AENA).

***Index Terms*—Air traffic control, Speech recognition, Command and control**

I. Introduction

CURRENT speech-based interfaces face the challenge of achieving acceptability in field applications, although a large degree of success has been obtained in specific areas such as medical or legal dictation. The main reason for the success in these environments is that the vocabulary is limited and specific, including long and easily identifiable words which enable high recognition test rates. The structure of sentences is mostly regular, requiring language models with lower complexity.

Both dictation and command & control systems have to compete with other well-established traditional interfaces such as keyboard and mouse. In some experiments carried out in a dictation task [3], errors with the keyboard were easy to identify just looking at the result on screen and took an average of 3 seconds to be corrected using additional key strokes. Alternatively, automatic speech recognition (ASR) errors were more difficult to locate and the average time for their correction was 25 seconds. However, for medical dictation applications, the keyboard is a less viable alternative [1] since physicians are used to recording machines and transcription services. Traditional manual procedures are slightly more accurate at the cost of a much higher turnaround time for the written report [2], compared to ASR. The harsh to admit higher error rates are counteracted by producing an overall decrease in medical costs and a faster service. In the case of legal services, dictation competes efficiently with the keyboard because there are many macros and shortcuts that can be used in a fast and easy way.

Other areas of application have emerged with the mobile concept. At the beginning of the use of mobile phones, speech recognition was only feasible as “distributed speech recognition” (DSR) whose idea was to perform all the CPU intensive calculations in central mainframes while only the speech capture and feature extraction ran in the mobile terminal. Wireless networks presented several problems that posed challenges to the performance of ASR systems: bandwidth constraints and transmission errors [4]. Recently the terminals have gained enough CPU power so that the idea of distributed recognition has declined and some laboratories are even introducing complex stochastic predictive algorithms in the terminal to reduce error rates [5]. For example, in [6] a 26% error rate reduction is achieved through learning language models from a large population for mobile command and control and a significant additional improvement of 5% through the online adaptation of the model to user specific data.

One of the sectors of the population that is more prone to accept speech interfaces is that of the persons with specific disabilities like persons that cannot use a keyboard or those with seeing impairments. In [15] an interesting study about ergonomics is presented that shows that about 75% of users with some of these disabilities continue to use ASR systems

J.M.Pardo, F. Fernández, J. Ferreiros, R. de Córdoba, J. M. Montero, R. San Segundo, L. F. d´Haro are with the Speech Technology Group (GTH), ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid Spain, Corresponding author J. M. Pardo, (pardo@die.upm.es, phone: +34 91 3367311, fax: +34 91 3367323).

J. Macias-Guarasa was with GTH when he contributed to this work and is now with Department of Electronics, University of Alcalá. Spain, current address Escuela Politécnica, Campus Universitario.Ctra.Madrid-Barcelona, Km. 33,600

V.Sama was with GTH when he participated in this work and is now with UNIDIS, UNED University, Madrid, current address C/ Fuente de Lima, 22. 28034 Madrid (Spain)

G. González is with ISDEFE (on behalf of AENA) current address: Edison 4, 28006 Madrid

This work was financed by a contract between AENA and UPM

and are satisfied with them. It is also shown that from the potential 150 words per minute (wpm) – promised in commercials for the speech product - a range from 8 to 30 wpm is achieved for people with these disabilities, highlighting the difference between laboratory and field performance measurements. Its conclusion about the learning curve for speech interfacing is also interesting. The study mentions that continuous speech recognition systems can be used with some success after only 2 hours of training, but the operating skill is still being developed after 20 hours of use over a period of weeks. In contrast, in the application presented in this paper, we use a very simple interface that shows high performance after short learning intervals (45 minutes) .

An important issue for a speech interface is whether it works for non-native speakers. This is our case in which controllers are typically language-native in Spanish that have to speak both in Spanish and in English at work. Even when using a single language, we have to consider dialectal variations as another source of difficulty. For Spanish native speakers (as in our experiments) the designer has to consider proper dialectal variations common in Spanish and also uttering variations produced when speaking in English able to model the higher uttering variability due to the lower proficiency in this language as compared to the mother tongue. It is also observed that some words (such as alpha, bravo, charlie, etc., city names, company names, etc.) may not be pronounced differently when they speak in English or Spanish and this contamination effect lead us to consider all the possible variants for both Spanish and English recognizers. This problem inevitability increases the test error rates and we had to introduce specific solutions similar to those presented in [7], i.e. introducing pronunciation variants for some words. This technique is not simple, because the introduction of pronunciation variants increases the recognizer perplexity and it has to be done carefully to obtain the desired gain in recognition.

Another problem of speech interfaces is the contamination of the speech signal with noises and the speech transmitting channel variability in a field environment. The solution to these issues implies more care in the design of the feature extractor. In [8] it is shown that for applications in which the signal to noise ratio (SNR) is greater than 15 dB, there are two key elements to take into consideration: an optimal setting of the feature extraction (in our case we study Cepstral Mean Normalization and Cepstral Variance Normalization techniques) and a proper setting for the “silence or non-speech” model (we use 14 different non-speech models).

In the history of the evaluation of the ergonomy of speech interfaces we find very positive opinions reached by Poock in the early 1980s [16], whose experiments showed a “very significant superiority” for speech over keying in a command and control application. Later Damper et al. showed that Poock´s experiments were carried out with an interface design more tailored to speech while easy improvements for the keying procedure could also be made. They carried out new experiments in an attempt to obtain a better experimental balance improving the keying interface and found that the big performance differences disappeared, although they reported that for cases where the user had to perform additional activities, many of the parallel tasks could be completed with the speech interface. Damper et al. stated in [17]: “Speech has an input potential for the future – especially for high workload situations involving concurrent tasks - if the technology can be developed to the point where most errors are attributable to the speaker rather than to the recognizer”. The application in this paper tries to fulfill the requirements set out in this statement as long as we are using the speech interface in an application were the user – the controller - has to perform many duties and we try to ease the interface with the systems commanded by the controller. It is well known that open and general speech recognition algorithms cannot attribute all the errors to the user, but we will show that by tailoring the vocabulary and controlling other aspects of the interface, test error rates can be reduced to usable figures.

There have been several other works connecting speech technology and ATC as in [9], where the authors, under the direction of the Avionics Engineering Center at Ohio University and supported by the FAA and NASA conducted tests on the idea of controller-pilot data link communications system. They used a Verbex speech recognition engine that achieved a 97% accuracy in a different task. Unfortunately, no performance figures were given for the proposed application, although they discussed the need for error correction and prevention procedures in order to consider the system usable. In [10] and [11] our research group presented results in a very similar task for Madrid Barajas Airport tower controllers, including the understanding of the commands for the controller-to-pilot communications channel. In [12], Duke et al took on a very big challenge: to develop an unmanned aircraft control system capable of operating in the national airspace system in Italy under instrument flight rules by using the voice communication channel and passing the Turing test in respect both to the conventional air traffic control and to other pilots in the area, as far as its answers and behavior was indistinguishable from other planes piloted by humans.

Other interesting applications are the use of speech technology for controller training, with the use of automatic pseudo pilots [13] and the estimation of controller workload [14]. These applications are examples of the quest for ideas where speech technology, with all its current weakness, may be really useful.

In our paper we describe the development, laboratory tests and field tests of a command and control speech interface of ATC terminals. We place the emphasis in the comparison between the laboratory and field experiments where a set of ergonomic-related effects are detected that cannot be observed in the laboratory experiments.

The paper is organized as follows: Section II describes the characteristics and features of the developed interface; Section III presents the hardware and software architecture; Section IV contains the laboratory development and tests; Section V shows the field objective and subjective evaluations and Sections VI and VII establish the discussion and conclusions of the work.

II. APPLICATION DESCRIPTION

AENA is the Spanish company in charge of Spanish airports and air navigation. They led the FOCUCS (Future control position, Sector Control Position unit) project in complete cooperation with the SACTA initiative (Automation system of ATC). FOCUCS is currently fully deployed in several ATC facilities.

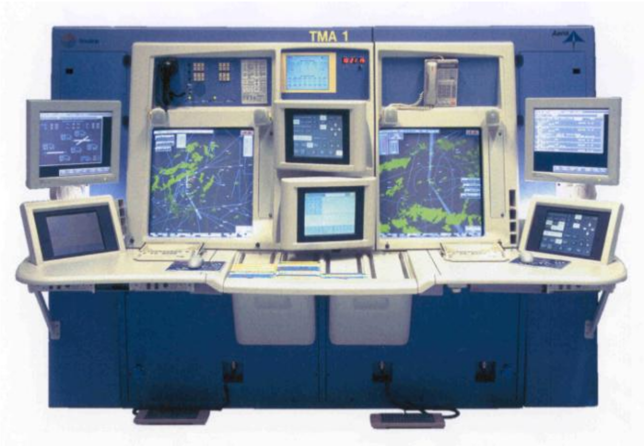

Figure 1 SACTA FOCUCS terminal

A FOCUCS terminal is made up of two sets for a pair of controllers working in a specific sector: one, the planning controller and the other, the executive controller (Figure 1). It consists of two main high resolution screens plus several other lateral screens, two of them touch-sensitive and other radio, audio and complementary equipment. The interface to the main screens in which the ATC information displayed is based on the keyboard and the mouse as well as a touch screen. The different functions and commands to get the information that appears on the main screens are introduced using two methods: with pop-up menus actuated by the keyboard or mouse and through strokes on virtual keys on the touch sensitive screen (for a limited list of relevant commands). For the case of pop-up menus, some commands appear under several (2 or 3) levels of selection.

The objective of the research contract between AENA and UPM was to explore the suitability of a speech interface to access the whole spectrum of commands via voice commands and see if this new interface could be faster and would alleviate the demands of handling the keyboard and mouse. As the controller's task is speech intensive, a special push-to-talk device had to be included to access the speech commander.

The task is not one of the most difficult ones for state-of-the-art speech recognizers in terms of vocabulary because it is very limited, but robustness was required in terms of the use of the system in real acoustics settings and professional environments. For the scalability of the system we decided that the vocabulary should be easily editable and expandable in the future. These characteristics forced us to consider different technological solutions as we will show later in this paper.

III. SYSTEM COMPONENTS

The hardware architecture of the system is presented in Figure 2.

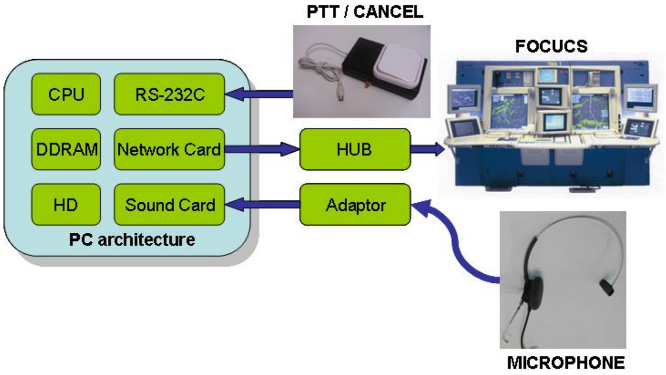

Figure 2 Hardware architecture of the system

The system runs in a personal computer and interacts with the FOCUCS terminal via a standard RJ45 network connection to a FOCUCS-local HUB. The speech is captured with an ATC standard microphone which is connected to an adaptor, developed within the scope of the project, to transform the differential balanced signal from the microphone to an asymmetrical signal that the PC sound card needs. Another component is a mouse-sized interface with two keys: one is a large push-button that will act as a PTT (Push to talk) and the other is a smaller lateral red push button that is intended to be actuated with the thumb and that will act as a canceller of the last command in the event of a human or recognizer error.

The software architecture that allows the operation of the system is set out in Figure 3. The system runs on Microsoft Windows XP operating system, making use of 3 application interfaces (APIs). The Winsock API is needed to be able to send the IP-UDP commands to the FOCUCS HUB. The multimedia API is needed for the management of the sound card, i.e., to sample the speech signal coming from the microphone adaptor. The PC port's API is needed to be able to read the PTT / Cancel push buttons interface. Above this API layer, we run three processes: the PTT / Cancel signaling module, the UDP client that will order the delivery of the right commands and our speech recognition engine. The commands understood by the system and other development information are presented via a Windows user interface on the PC.

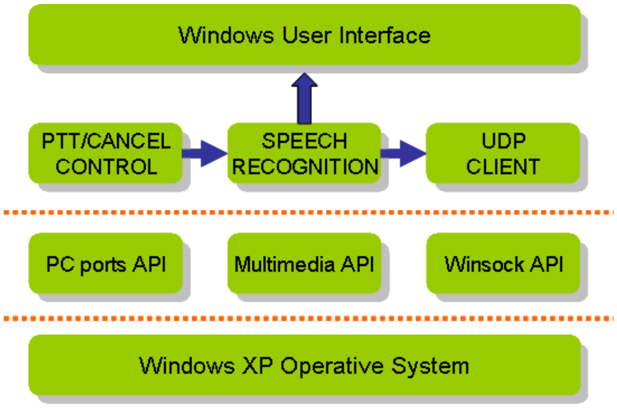

Figure 3 Software architecture of the system

The recognizer itself is made up of the main modules presented in Figure 4. The first module calculates the end of the command utterance (Note that the start is given by the action of the user on the PTT). This module compares the energy of the signal to three energy thresholds which are all relative to a floor energy estimation for silent segments. For any candidate end point, the energy must remain below the medium energy threshold and the low energy threshold for a pre-defined time. If after some second pre-defined time the energy rises again and crosses the highest threshold level, the event is considered a pause between two words (which happens in some compound commands) and the decision process is reset until the new word ends. The time and energy threshold factors can be adjusted.

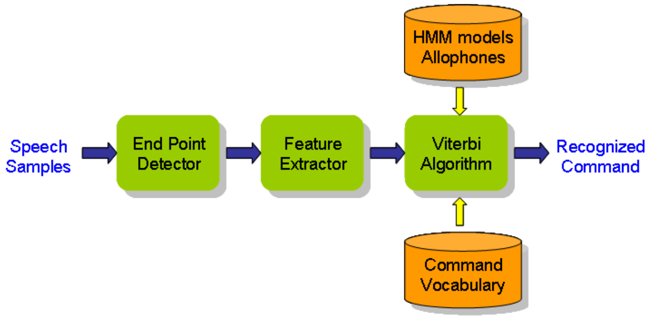

Figure 4 Recognizer mechanism

Once we have segmented the command utterance, the speech samples enter the feature extractor. The recognizer uses 13 LPC-Cepstrum [18] base features each of 10 msec. temporal analysis window (25 msec. wide each). These features are processed using CMN (Cepstral Mean Normalization) and CVN (Cestral Variance Normalization) in order to minimize the effect of the acoustic propagation channel of the speech signal. In other words, CMN and CVN achieve a certain degree of robustness against changes in microphones, in lines and in pre-amplifiers through which the signal may be transmitted before it enters our system. We also take the 13 first and second derivatives of these normalized LPC-Cepstrum features to compose a final feature vector of 39 components.

The speech recognizer is based on the Viterbi algorithm [19] that calculates a score between the spoken utterance and the models for all the commands within the specified vocabulary, selecting the best scored one. The models are stochastic continuous density Hidden Markov Models (HMMs) [20] for allophones (as we will see later, we use context-dependent allophones). The recognizer builds up the model for each command to be recognized (and even for each possible pronunciation variant of the same command) concatenating the corresponding allophonic sequence models read from the vocabulary. This characteristic of the recognizer allows the easy expansion of the vocabulary of terms to be recognized (or the pronunciation variants of existing commands), by just writing down new entries in the command vocabulary.

IV. Laboratory Development and Evaluation

A. Database

To build up the recognizer stochastic models, a large speech database is needed. We selected the SpeechDat database[1] (the isolated words part with 41.8 hours of speech) for the training of the Spanish recognizers because it matched our requirements very well for this project. The main characteristics of SpeechDat are as follows:

- It consists of 4,000 speakers
- It has been recorded through a narrow band (4KHz) telephone speech channel, very similar to the ATC equipment channel
- Several tasks are recorded, including commands

To complement this database, we recorded another specific database in the context of the project that we call the Invoca database. This database consists of recordings from 27 speakers (14 males and 13 females) from 18 to 40 years old. Each speaker uttered 5 repetitions of each command (228 words for Spanish and 146 words for English). Two subsets were created: 18 speakers (20,380 files containing 10.39 hours of speech) were kept both for task adaptation of the models and for testing the alternative of training from scratch; the other 9 speakers (10,220 files) were separated for speaker adaptation techniques tests and for the final laboratory validation of the system.

We decided to use context-dependent models at the tri-allophone level in which the states that make up a model can be shared with other similar states. The tri-allophone models are allophone models that are different depending on the lateral allophones adjacent to the considered center allophone. For example, when modeling an allophone [p],

[1] http://catalog.elra.info/product_info.php?products_id=635, as of september 2009.

instead of having just one model for all [p] occurrences we have different models if the [p] is preceded by a certain allophone [p_1] and followed by another allophone [p_2], thus constituting the allophone [p_1 p p_2]. We have to keep in mind that [p_1 p p_2] is a model just for the sound [p], not for the sequence “p_1 p p_2” but only for the occurrences where “p” is preceded by the allophone p_1 and followed by p_2. If we would consider all possibilities for the contexts, the modeling would be too big and not trainable. This is the reason to perform states sharing among all possible states resulting from the use of the tri-allophone idea. The objective of sharing states in a cluster is to obtain robust states. We mean robust in the sense that all states have enough training material for a reliable estimation of the model parameters. The clustering procedure (estimation of groups of “similar” or “close” states) is based on a phonetic decision tree, which is a binary tree in which a yes/no phonetic question is attached to each node. Initially all states in a given item list (typically a specific phone state position) are placed at the root node of a tree. Depending on each answer, the pool of states is successively split and this procedure continues until the states have trickled down to leaf-nodes. All states in the same leaf node are then tied. The question at each node is chosen to (locally) maximize the likelihood of the training data given the final set of state tying. Each question takes the form “Is the left or right context in the set P?” where the context is the model context as defined by its logical name. More details can be found at [21].

We will now summarize the process of the development of the Spanish recognizer with which the main design decisions were made.

B. *Development of the system*

Our initial HMM models were estimated applying the Baum-Welch algorithm [19] to the SpeechDat (isolated words part) database. With these initial experiments we found the optimum regression tree in 1,509 states with a mixture of 6 gaussians per state and we optimized several parameters of our system. With these models, we obtained 2.08% test error rate for the ATC terminal command task. It is important to note that the SpeechDat database does not include any of the ATC terminal commands, so some allophonic contexts needed for the task could not exist in the SpeechDat training database. However the SpeechDat database has a lot of general training material so that our recognizer benefits from the generality of the HMM models obtained. This feature is relevant in order to augment the commands in the future.

Our next step was to test whether using task adaptation would improve our results. Task adaptation means the use of some data from the proposed task to modify (adapt) the models obtained with speech material coming from another task in order to improve the performance. This is clearly our case in which we have SpeechDat-based trained HMM models whilst we would like the best performance on our ATC terminal command task. Thus, we used the Invoca subset of 18 speakers to adapt the HMM models and then we tested the results on the subset of 9 speakers. The results appear in Table 1, column 2 where two versions of two different adaptation techniques are presented.

		Speaker adaptation		
	Task adaptation	1 rep	2 rep	3 rep
MAP v1	1.00%	0.29%	0.17%	0.17%
MAP v2	0.81%	0.27%	0.17%	0.17%
MLLR v1	0.84%	0.47%	0.39%	0.29%
MLLR v2	0.84%	0.27%	0.27%	0.15%

Table 1. TEST WORD ERROR RATES FOR EXPERIMENTS OF MODEL ADAPTATION: THE FIRST COLUMN SHOWS TASK ADAPTATION AND THE 3RD TO 5TH COLUMS, THE SPEAKER ADAPTATION EXPERIMENTS

MAP (Maximum a posteriori) adaptation [22] uses the original model to estimate the probability of a representation link between each acoustic frame and the model states and then estimates a new model as an interpolation of the original parameters and those extracted from the new data, considering this estimate of linking. In Table 1 “MAP v1” stands for the use of CVN-compensated features and adaptation of only the means of the gaussians via MAP, and “MAP v2” stands for the CVN compensation of both means and variances with MAP. In the MLLR (Maximum Likelihood Linear Regression) technique [23][24], the adaptation data is used to learn simple regression trajectories for the gaussian parameters, in our case just for the means in the “MLLR v1” version and of both means and variances in “MLLR v2”. The best result is found for MAP v2 with a 0.81% test error rate.

This result was considered enough to start the field evaluation (as we will see later), but we checked another adaptation possibility that consisted of using data from a specific speaker to further improve her/his model, thus adapting the previous models (speaker-independent models) to the personal pronunciation particularities of the user. For this new experiment we used the 9 speaker sub-set of the Invoca database and we started from the task-adapted models obtained from the previous phase. We decided to use two of the repetitions for testing in all the cases and use the remaining three repetitions to carry out different experiments. Three experiments were made: adaptation with just one repetition, with two repetitions or with the three repetitions (1.2 and 3 repetitions had an average of 7.1, 13.8 and 20.8 minutes of speech respectively). The results are presented in columns 3, 4 and 5 of Table 1. Obviously, the system performs better when more data is used to adapt the models. We were also interested in evaluating how much adaptation data we would need for the addition of a new controller to the system. Again the best technique option is MAP v2 which produces a reduction to a 0.27% test error rate with one repetition of the commands, to 0.17% with two and three repetitions. One repetition of all the commands means the repetition of 228 words in Spanish, a list that could be long. In a subsequent experiment we calculated the results of using the adaptation algorithm with a shorter list of words. We selected 50 words among the most acoustically confusing words (the ones that produced more errors in previous experiments) and at the same time more representative of the task (they better cover the original distribution of the allophones). These 50 words had an average duration of 1.6 minutes in total. The results of the experiment adapting with the short list are presented in Table 2.

	Speaker adaptation with the 50 words list
MAP v1	0.56%
MAP v2	0.56%
MLLR v1	0.61%
MLLR v2	0.54%

Table 2. TEST WORD ERROR RATES FOR EXPERIMENTS OF MODEL ADAPTATION WITH THE 50 WORDS LIST

The results are in between the previous results (i.e. no speaker adaptation or full speaker adaptation). However the MAP technique is less effective in this case (the test error rate doubles from 0.27 % to 0.56%) while MLLR v1 performs 0.54%, which is the first experiment in which MLLR gives a lower test error rate than MAP.

It is also true that with the confidence margins produced by the low number of test words of the experiments, the differences between all speaker adaptation techniques were not statistically significant. This is also the reason for not carrying out speaker adaptation experiments in the English command recognizer as we even had less data for this task. For English we created the system by directly training the models with the Invoca database and it resulted in a 2.7% word error with the 241 word list. This performance was considered adequate for the purposes of this project. The optimum in English was found for 1,400 states and a mixture of 8 gaussians per state.

We finally decided that the improvement obtained by speaker adaptation was not big enough to compensate the discomfort of using an enrollment procedure for every new speaker (who has to record the set of adaptation words) and the field experiments were carried out with the models adapted to the task but without speaker adaptation.

V. FIELD EVALUATION

The whole system was evaluated both in English and in Spanish. The purpose of the evaluation was twofold: we wanted to know actual field test error rates and we also wanted to extract information on the ergonomics of the new interface and to know about the opinion on usability of the users.

The evaluation consisted of two phases, one called "guided evaluation" where the users had to utter specific items as they appeared on a screen. In this way, the system could automatically calculate the error rate. The other phase, called "free evaluation" was mainly intended for ergonomics evaluation and consisted of realistic working scenarios in which the interface was used to solve the programmed task. An external observer annotated the number of recognition errors in this second phase so that we also have this figure, but the main outcome was obtained through the answers from the users to a questionnaire with several questions about the usability of the interface. For these experiments we involved 22 users: 11 volunteers from the AENA Automation Division that made the guided evaluation and 11 controller students from SENASA (the controller's training facility) who made the free evaluation on realistic scenarios.

A. Guided evaluation

The system was prepared so as to recognize 412 variants (different words plus pronunciation variants) in Spanish for a total of 228 commands and 383 variants in English for a total of 118 commands. A command may imply the uttering of several words and each word may have several alternative pronunciations, mainly for English as the speech is coming from non-native users. The 11 speakers uttered a list of 50 commands in Spanish and 30 in English. We separated the results in two sets: "single commands" is the one where each command consisted of just one word and "compound commands" is the one where each command consisted of several words and all of them have to be correctly recognized to consider the command recognized. The results that we have obtained are summarized in Table 3. The line labeled "Overall" integrates the results for single and compound commands

SPANISH	word error	command error
single commands	2.6%	2.6%
compound commands	3.5%	7.3%
Overall	3.3%	6.0%

ENGLISH	word error	command error
single commands	4.9%	4.9%
compound commands	5.4%	12.4%
Overall	5.3%	10.6%

Table 3. RECOGNITION ACCURACY FOR FIELD "GUIDED EVALUATION"

Single commands perform better than compound commands because the latter need more than one word to be correctly recognized. For the same reason, compound commands evaluated at the word level give better results than the full-command recognition test rate. Overall results integrate both single and compound commands so their results are the reference results.

The errors of the field tests are considerable higher than the ones obtained in the development phase. An error analysis could be made since we had all the utterances recorded and labeled. The result of this analysis is shown in Table 4.

	Spanish	**English**
Empty file	1.5%	1.5%
Cut off recording	0.4%	0.5%
Repeated word	0.1%	0.0%
Badly labeled	0.9%	0.0%
Wrong pronunciation	0.0%	0.3%
Noise	0.0%	1.5%
Rest of errors	0.4%	1.5%
TOTAL	**3.3%**	**5.3%**

Table 4. FIELD RECOGNITION ERRORS ANALYSIS

The PTT procedure and the inexperience of the subjects of this experiment in the use of speech recognizers (they were volunteers from the Automation Division and not student controllers who at least have more experience in using PTT procedures) caused most of the system problems: some files were found to be empty of any signal or with a partial recording of the command with a significant part cut off (rows 1 and 2 of Table 4). In some cases we found the word uttered twice within the analysis window. In some other cases we found an error in our labeling procedure when writing down the reference text that caused the command not to be counted as recognized. Other cases contained errors in the pronunciation of the command (in English) or were corrupted with a significant noise. If we do not consider these cases, we find 0.4% words for Spanish and 1.5% for English to which we cannot impute any significant problem in the recording and these figures correlate quite well with the laboratory expectations for this "controlled" field guided evaluation.

The results obtained are the natural consequence of using a system prepared in the laboratory in the field by real users and the "realistic recognition test error rate" is 6% for Spanish or 10.6% for the English overall command test error rate that includes the effect of all real field application problems and details.

B. *Free evaluation*

AENA´s DOR (Organization and Ruling Division) designed some realistic protocols or scenarios for interaction with FOCUCS. We asked the 11 student controllers to complete them using the speech interface instead of the traditional access through keyboard and mouse or touch commands. They were free to perform the task in any order and using the commands they would consider necessary at each moment. We gave them a briefing of about 45 minutes explaining the capabilities and operation of the new speech interface. They had to use both Spanish and English commands in these scenarios.

After the completion of their task, they had to fill in a questionnaire with several questions. While they were working, an external observer manually wrote down the number of recognition errors that came to 65 out of 1,157 observed commands. That makes a 5.62% command test error rate in this field use of the system, which is better than the "guided evaluation" results of previous section. The reason for this improvement can be attributed to the greater experience of the student controllers in the use of PTT procedures and the familiarity with the commands that produce better and more consistent pronunciations.

As SENASA gives controller formation to students coming from all the regions in Spain, there are several pronunciation variants within the testing subjects. The distribution of the subjects pronunciation regions are shown in Figure 5.

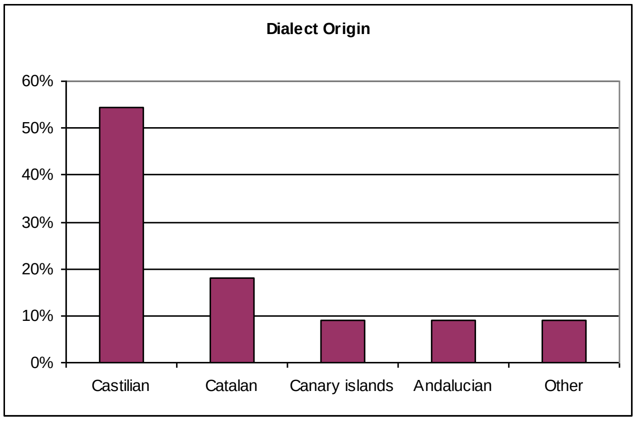

Figure 5 Dialectal origin of "free evaluation" population.

The main objective of this test was to learn about the usability of the speech interface as experienced by the testing subjects. Their feeling was captured through the answers to a questionnaire with the questions that appear in the head of each sub figure in Figure 6.

The first question was about prior experience in speech recognition (Q1). The results show that most of the users were not used to speech interfaces, even though they are enrolled in a high technology-related training like that of an ATC controller. This factor is something that speech technology has to fight against: the majority of the population has a high degree of inexperience in the use and particularities of human-machine interfacing with speech.

The second question (Q2) is whether the system is easy to use. The results indicate that the system is perceived to be easy to use, validating our efforts to take the user into account when we developed the system.

The two following questions (Q3 & Q4) about the understanding capabilities of the system and the speed of execution are also answered quite positively with more than 82% of the users happy with the new interface (Agree + Fully agree for both questions).

In Q5, we asked whether they found the phraseology adequate. 81% of the answers considered it adequate or very adequate, although we find that 9% of the subjects did not like it. The phraseology was prepared by AENA Automation Division in collaboration with the development team and this question clearly indicates that we have to consider user suggestions for future versions as they are the people that will actually use the interface and know about the phraseology preferences and operability of the commands.

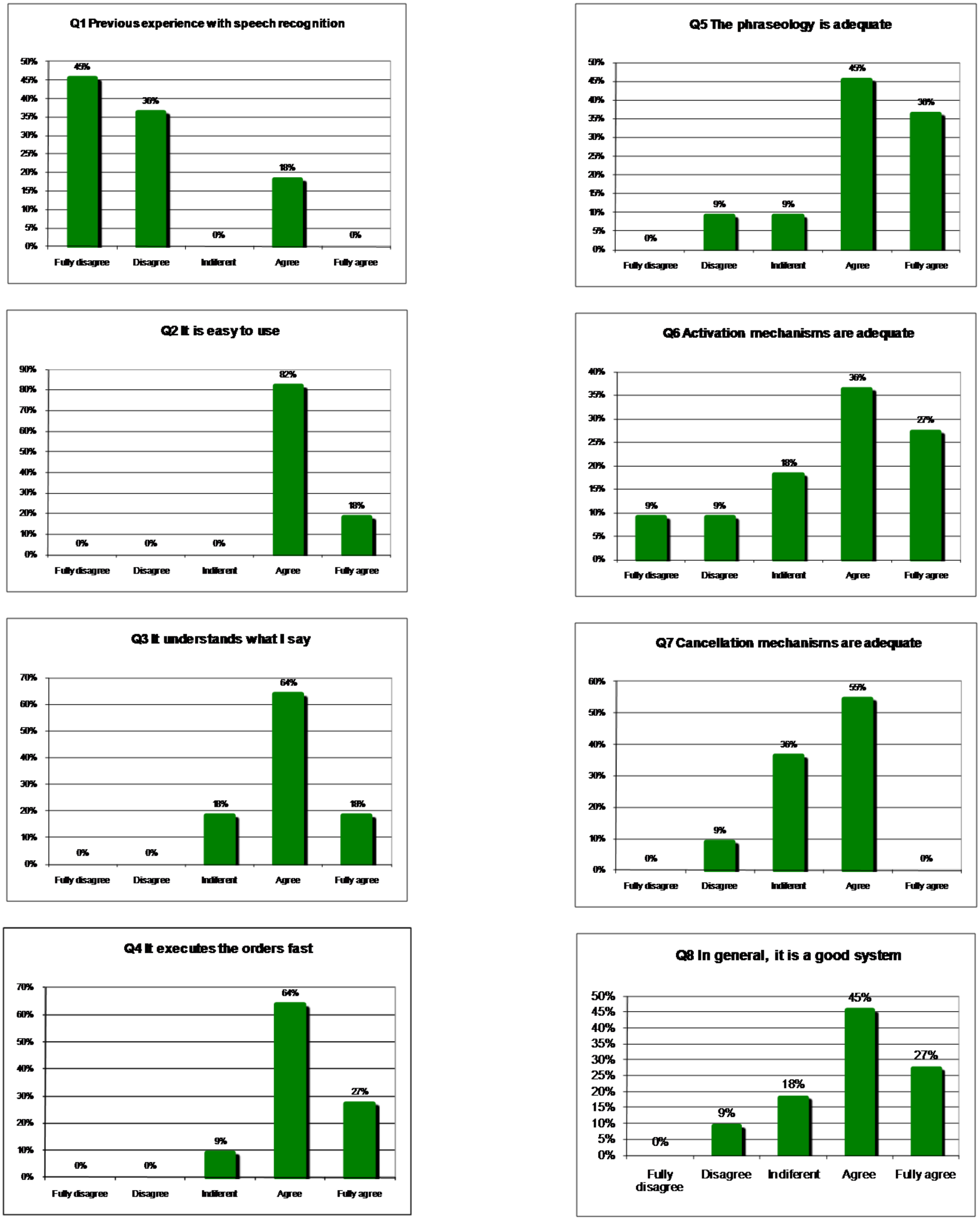

Figure 6 Answers to the questionnaire

The following question (Q6) was whether the activation mechanism (PTT) was adequate. Although the results are quite positive, the answers also point out another weakness in the interface with 18% of the subjects with moderate or severe trouble when using this interface. We can also connect this question with the following one (Q7) about the cancellation mechanism with similar although not extreme answers. Our conclusion is related to the PTT procedure. Controllers have to use the radio PTT for their duty and we are forcing them to use a second different PTT for the recognizer and a special key near this PTT for the cancellation of the last command in the event of misrecognition. This is not the best way to achieve a good ergonomy and we have to consider a better way to integrate the interface for this application in the future. One idea could be to use a modified PTT that would integrate both the PTT for communications and that for recognition and the cancellation key.

The last question (Q8) summarizes the user feeling by asking whether the system is a good system. The result is quite positive: 72% are happy, 18% indifferent and only 9% did not like the system, with no one voting for a full disagreement with the statement of the question. This fact encourages us to consider user suggestions to try to improve user acceptability.

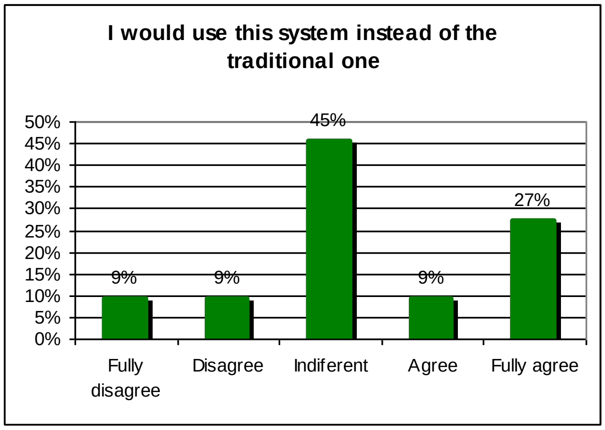

Figure 7 Answer to the most relevant question: would the user use the speech-based system instead of the traditional one?

Finally, we end with the answers to the relevant sentence “I would use this system instead of the traditional one”, an overall question about the usability of the new interface (see Figure 7). The answers are broken down into 36% positive, 45% indifferent and 18% negative, not a bad result for an interface competing with well-known and extensively used interfaces as is the case of keyboard and mouse plus the help of the touch screen for some commands. Our users also have the challenge to find an interface that makes use of the same intellectual channel (speech) that they use for the communications with the pilots or collaterals on their duty. This is the reason for considering the answers again encouraging.

VI. Summary & Discussion

Table 5 summarizes the speech recognition results for the different tests of the system.

Experiment	Spanish	English	Comments
Laboratory	0.8%	2.7%	word error rate
Field, guided evaluation	3.3%	5.3%	overall word error rate, all effects included
	0.4%	1.5%	overall rest word error rate
	6.0%	10.6%	overall **command** error rate, all effects included
Field, free evaluation	5.6%		overall **command** error rate, all effects included

Table 5. Test recognition errors summary

In the first row we present the results of the development in the laboratory, 0.8% word error (Spanish) and 2.7% word error (English).. When we carried out the field tests, first with the guided evaluation, where the speakers had to utter specific commands as they were required by the automatic evaluation system, these test word error rates rapidly rise to 3.3% and 5.3%. After a close look at the causes of this increase in error, we find the effects that we have explained before when real users confront the interface. These effects include empty files, others with just part of the utterance, repetitions of the word within the analysis window, wrong pronunciation in English and inadmissible noises. Most of these “extra” effects come from the way of using the interface and should be dealt with using a thorough study of the best ways to produce a match between system requirements and real user use of the interface. If these effects could be minimized, we could come close to the laboratory performance (in our test, the estimation is of 0.4% and 1.5% as can be seen in the third row of Table 5). Considering that our task needs the full command to be recognized even for those cases where the command consists of several words, we look at the fourth row and find a performance of 6% (Spanish) and 10.6% (English) full test command error rate. These errors will have to be corrected by making use of our “cancel” button prepared in the new interface, a feature that we consider eventually necessary for all speech interfaces as a 0% error rate is unfeasible in the laboratory and certainly in real working environments. This additional cancellation procedure has also been evaluated in the usability subjective tests.

When we analyze the free evaluation tests, we observe a 5.6% test command error rate, where Spanish and English commands are mixed up freely in the carrying out of the scenarios. In spite of this effect, we notice a better performance than in the guided evaluation. We must remember that the reason is that the free evaluation is made by student controllers that are more familiar with both the terminology and the PTT procedures.

On the ergonomics and usability side, the answers to the questionnaire show the following summarized tendencies in opinion: First, we see a low degree of previous experience on the use of speech technology products, the same as for the general population to the best of our knowledge. With this starting point of low experience of the users, they nevertheless consider our interface easy to use, they consider that the system actually understands their commands and they believe that the system reacts rapidly to their demands. When they give their opinion on the phraseology, although positive again in general, we find 9% of the answers negative which makes us think that more discussion with the final users is necessary to refine the vocabulary for the commands for the eventual use of the system in actual production. The activation mechanism (PTT for recognition) is received somewhat negatively with 18% of the answers revealing moderate or severe trouble with this procedure. As we have already mentioned, we have analyzed the two main reasons for this problem: Our interface competes with other communications PTT, physically apart from each other and we are using the same intellectual channel (speech) that may produce some confusion to the user as he has to think to which of the receivers his/her commanding utterances are aiming at. The cancellation procedure shares the same tendency of opinion but a little attenuated and we think that this may relate to the fact that the users see the benefits of fast recovery from error that may come from the recognizer or from a mistake in the order. The strongest usability question in the test is posed by the answers to the statement “I would use this system instead of the traditional one” and we find the distribution of 36% positive, 45% indifferent and 18% negative opinion that encourages us to carry out further studies to bring this interface closer to actual production once the aforementioned difficulties are overcome.

The sequenced project workload is given in Table 6 in order to perceive the intensity of each constituent task. Most of the effort was needed to prepare and adapt the speech recognizer (about 40% of the total effort). Another important part of the effort is used in the analysis of the control panel and the subsequent phraseology set up both for Spanish and English (about 30% of the total). We discovered that this part is critical and needs good synergy with the users in order to obtain a usable speech interface. Two issues were balanced in these phraseology preparation tasks: the designers heard the suggestions of the users for the commands and modulated their wills by the expertise knowledge about the acoustic confusability of the proposed commands. A consensus was reached to define ergonomic commands that exhibit the lower acoustic confusability possible. This is a relevant and time consuming task that must be carried out with care to pursue the highest success possible. Finally, we also note that about 18% of the effort was dedicated to evaluations and analysis that have also to be carefully performed to obtain sensible conclusions.

Control panel analysis	3 p·m
Phraseology design and validation	8 p·m
Specifications of phraseology for English version	2 p·m
Works on the recognition engine	18 p·m
Tools to define command to action connections	2 p·m
User interface design	3 p·m
Laboratory evaluations	3 p·m
Field prototype evaluations	4 p·m
Results analysis	1 p·m
TOTAL	**44 p·m**

Table 6. SEQUENCED PROJECT WORKLOAD IN PERSONS·MONTH (p·m) BY TASKS

VII. CONCLUSIONS

We have designed an automatic speech recognition command and control system for ATC terminals, and we have made a field evaluation of the recognizer and of the full system by analyzing test command error rates as well as the ergonomics and usability of the interface. Speech recognition test error rates are low and this speech recognizer was judged usable by the users although the need for a canceling button for the cases of recognition or user error is also clear. The phraseology needs further refinement to reach full user acceptability, but the subjective evaluation shows that the users are already prepared to use this interface even when it is compared to the traditional and widely accepted keyboard and the mouse or even the touch input. Field performance of the recognizer worsens as a result of field effects not seen in the laboratory tests, making field tests and their analysis mandatory for the final design of the system and the eventual success of the interface.

ACKNOWLEDGMENT

We express our gratitude to the personnel in AENA and SENASA who helped us in many stages of the development of this project, especially to Myriam Santamaría. We also wish to thank J. Colás for his work in the first stages of the project.

REFERENCES

[1] Giovanni Mazza, Marcie Battles, and Helmuth F Orthner, “Application of Speech Recognition Data Entry Enhancements in an Electronic Patient Care Report (ePCR)”, *AMIA Annu Symp Proc. 2006*.

[2] Grasso, M. A. (2003). The Long-Term Adoption of Speech Recognition in Medical Applications. 16th IEEE Symposium on Computer-Based Medical Systems (CBMS 2003), 257-262. Retrieved from http://citeseer.ist.psu.edu/626752.html

[3] Koester H. H. (2001). User performance with speech recognition: a literature review”.Assist Assistive Technology 13(2):116-130.

[4] Z.-H.Tan, P.Dalsgaard and B.Lindberg, “Automatic speech recognition over error-prone wireless networks”, Speech Communication 47 (2005) 220–242

[5] Andreas Hagen , Daniel A. Connors, Bryan L. Pellom , “The analysis and design of architecture systems for speech recognition on modern handheld-computing devices”, Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, 2003

[6] Tim Paek, David Maxwell Chickering, “Improving command and control speech recognition on mobile devices: using predictive user models for language modeling”, User Model User-Adap Inter (2007) 17:93–117

[7] Judith Kessens, “Non-native Pronunciation Modeling in a Command & Control Recognition Task: A Comparison between Acoustic and Lexical Modeling”, ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006)

[8] Jan NOVOTNÝ, Pavel SOVKA, Jan UHLÍŘ, “Analysis and Optimization of Telephone Speech Command Recognition System Performance in Noisy Environment”, RADIOENGINEERING, VOL. 13, NO. 1, APRIL 2004

[9] Alicia Lechner, Patrick Mattson, Kevin Ecker, “Voice Recognition: Software Solutions in Real-time ATC Workstations”, IEEE aerospace and electronic systems magazine ISSN 0885-8985, 2002, vol. 17, no11, pp. 11-16

[10] F. Fernández, J.Ferreiros, J.M.Pardo, V. Sama, R. de Córdoba, J. Macías-Guarasa, J.M.Montero, R. San Segundo, L.F, d´Haro, M. Santamaría, G. González, “Automatic Understanding of ATC Speech”, IEEE A&E Systems Magazine, October 2006, pp 12-17.

[11] J. M. Pardo, J. Ferreiros, F. Fernández, V. Sama , R. de Córdoba, J. Macias-Guarasa, J. M. Montero, R. San Segundo, L. F. d´Haro, G. González, “Automatic Understanding of ATC Speech: Study of Prospectives and Field Experiments for Several Controller Positions”, accepted for publication in IEEE Transactions on Aerospace and Electronic Systems (TAES), to appear soon.

[12] E.L.Duke, C.C. Vanderpool, W.C, Duke,. “Turning PINOCCHIO into a real boy: Satisfying a turing test for UA operating in the NAS”, Collection of Technical Papers - 2007 AIAA InfoTech at Aerospace Conference, Volume 2, pp. 1675-1690.

[13] D. Schäfer “Context-Sensitive Speech Recognition in the Air Traffic Control Simulation,” in Universität Der Bundeswehr München Fakultät Für Luft- Und Raumfahrttechnik, Phd. Thesis,2001, and Eurocontrol Experimental Centre EEC Note No. 02/2001. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8234&rep=rep1&type=pdf

[14] Carol Manning, Scott Mills, Cynthia Fox, Elaine Pfleiderer, and Henry Mogilka, “The Relationship Between Air Traffic Control Communication Events and Measures of Controller Taskload and Workload”, 4th USA/Europe Air Traffic Management R&D Seminar. http://www.atmseminar.org/seminarContent/seminar4/papers/p_161_HF.pdf

[15] Heidi Horstmann Koester, “User performance with speech recognition: a literature review”, Asst Technol 2001; 13:116-130

[16] Poock, G.K. “Voice recognition boosts command terminal throughput”, Speech Technology, 1, 36–39, 1982

[17] R.I. Damper, M.A. Tranchant and S.M. Lewis “Speech versus Keying in Command and Control: Effect of Concurrent Tasking”, International Journal of Human-Computer Studies, 1995. http://eprints.ecs.soton.ac.uk/73/2/concurr.pdf

[18] H. Matsumoto and M. Moroto, “Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2001, vol. 1, pp. 117–120.

[19] A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. L.R. Rabiner. Proceedings of the IEEE, Vol. 77, n. 2, February 1989

[20] Spoken Language Processing, Huang, X., Acero, A., Hon, H.W. Ed. Prentice Hall, New Jersey, 2001

[21] “The HTK Book”, Steve Young et al. Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/.

[22] Gauvain, J.L., Lee, C.H., “Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, IEEE Trans. SAP, Vol. 2, pp. 291-298, 1994

[23] Gales, M.J.F., Woodland, P.C., “Mean and Variance Adaptation Within the MLLR Framework”, Computer Speech & Language, Vol. 10, pp. 249-264, 1996.

[24] Leggetter, C.J., Woodland, P.C., “Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression”, Proc. ARPA SLT Workshop, pp. 104-109. Morgan Kaufmann. 1995.

Javier Ferreiros (M’02) earned his Telecommunication Engineering Degree and PhD both from Universidad Politécnica de Madrid in 1990 and 1996 respectively. Since 1989 he has worked in Speech Technology and has held different teaching and research positions at the Universidad Politécnica de Madrid. He is Associate Professor since 2001. He was Associate Director of the Electronic Engineering Department from 2004-2008. and is currently Associate Director for Academic Planning of the Telecommunication Engineering School of the Universidad Politécnica de Madrid Prof Ferreiros was a visiting scientist at the International Computer Science Institute in 1999-2000. He was the Technical Program Manager of EUROSPEECH 1995. He is member of ISCA. He has authored or co-authored around 100 papers and holds one patent.

José M. Pardo (M’84-SM’04) earned his Telecommunication Engineering Degree and PhD both from Universidad Politécnica de Madrid in 1978 and 1981 respectively. He won a National Award in 1980 for the best graduate in Telecommunication engineering and a National Award for the Best PhD Thesis in 1982 . Since 1978 he has worked in Speech Technology and has held different teaching and research positions at the Universidad Politécnica de Madrid. He has been the Head of the Speech Technology Group since 1987 and Full Professor since 1992. He was head of the Electronic Engineering Department from 1995-2004. Prof Pardo was a Fulbright Scholar at MIT in 1983-84, a visiting scientist at SRI International in 1986 and recently a visiting fellow at the International Computer Science Institute in 2005-2006. He was member of the ISCA Advisory Council from 1996 until 2006. Prof. Pardo was chairman of EUROSPEECH 1995 and member of ELSNET Executive Board 1998-2004. He was member of NATO RSG 10 and IST 3 from 1994 to 2002. He is member of JASA, ISCA and EURASIP. He has authored or co-authored more than 160 papers and holds two patents.

...