# Automatic Understanding of ATC Speech: Study of Prospectives and Field Experiments for Several Controller Positions

**J. M. PARDO,** Senior Member, IEEE
**J. FERREIROS,** Senior Member, IEEE
**F. FERNÁNDEZ, V. SAMA**
**R. DE CÓRDOBA,** Member, IEEE
Universidad Politécnica de Madrid

**J. MACIAS-GUARASA,** Member, IEEE
University of Alcalá

**J. M. MONTERO,** Member, IEEE
**R. SAN-SEGUNDO, L. F. D'HARO**
Universidad Politécnica de Madrid

**G. GONZÁLEZ**
ISDEFE

**Although there has been a lot of interest in recognizing and understanding air traffic control (ATC) speech, none of the published works have obtained detailed field data results. We have developed a system able to identify the language spoken and recognize and understand sentences in both Spanish and English. We also present field results for several in-tower controller positions. To the best of our knowledge, this is the first time that field ATC speech (not simulated) is captured, processed, and analyzed. The use of stochastic grammars allows variations in the standard phraseology that appear in field data. The robust understanding algorithm developed has 95% concept accuracy from ATC text input. It also allows changes in the presentation order of the concepts and the correction of errors created by the speech recognition engine improving it by 17% and 25%, respectively, absolute in the percentage of fully correctly understood sentences for English and Spanish in relation to the percentages of fully correctly recognized sentences. The analysis of errors due to the spontaneity of the speech and its comparison to read speech is also carried out. A 96% word accuracy for read speech is reduced to 86% word accuracy for field ATC data for Spanish for the "clearances" task confirming that field data is needed to estimate the performance of a system. A literature review and a critical discussion on the possibilities of speech recognition and understanding technology applied to ATC speech are also given.**

Manuscript received December 16, 2008; revised December 12, 2009, and April 6 and June 29, 2010; released for publication September 8, 2010.

IEEE Log No. T-AES/47/4/942897.

Refereeing of this contribution was handled by M. Efe.

Authors' addresses: J. M. Pardo, J. Ferreiros, F. Fernández, R. de Córdoba, J. M. Montero, R. San Segundo, and L. F. d'Haro, ETSI de Telecomunicación, Ciudad Universitaria, 28040 Madrid, Spain, E-mail: (pardo@die.upm.es); V. Sama, UNIDIS—(UNED-Fundación MAPFRE) Centro de Atención a Universitarios con Discapacidad, C/Fuente de Lima, 22 28024 Madrid, Spain; J. Macias-Guarasa, Escuela Politécnica Superior, Departamento de Electrónica, Carretera Madrid-Barcelona, Km 33,600, C.P.28871, Alcalá de Henares, Madrid, Spain; G. González, Isdefe, S.A., c/Edison 4, 28006 Madrid, Spain.



## I. INTRODUCTION

Speech technology is the area of science that allows the processing of human speech for its recognition, understanding, translation, and generation. Current speech technology applications can be divided into online and offline applications. Online applications include the following: 1) dictation, in which a person speaks to a computer and the system transcribes what is spoken [1–3], 2) telephone-based applications, in which the computer interacts with the user by recognizing/understanding the question and generating a useful answer [4], 3) applications in a car or an environment in which the hands of the user are busy in which the system helps with user needs, for instance at home or a production factory or in the case in which there are no keyboards (i.e., a mobile phone) [5–7], and 4) Language learning in which the system acts as a tutor to the student [8–10]. Offline applications include processing recorded audio for indexing it and its subsequent rapid recovery or extracting information from it.[1] Speech technology applied to ATC speech can be used in both scenarios, online for an ATC training application or offline for the analysis of an ATC task load. The sophistication of the techniques used depends on the application.

Several researchers have tried to process ATC speech in the past for different purposes. The first reference to it that we found in the literature dates back to 1975 in which a limited speech understanding system was studied for use as a component in a military training system [11] mentioned by Beek, et al. [12]. In 1990, while discussing the potential of speech processing in military computer-based systems, Weinstein mentions the application of training air traffic controllers as a way of eliminating the need for a person to act as pseudopilot thus reducing the cost of training personnel [13]. Methods of training air traffic controllers include the use of human pseudopilots that mimic a working scenario. The controller interacts with the pseudopilot in the same way as he/she would interact when he/she is on duty. One of the problems of this methodology is the cost of training and paying the human pseudopilots. The idea of the ATC training simulator comprises the following: 1) the ATC speech is processed and understood by a speech understanding module of the pseudopilot system; 2) the central control of the automatic pseudopilot system includes a model of air traffic procedures (and possibly a model of the air traffic controller's behavior and performance modeling) which then generates a response that is sent back to the air traffic controller for the following interaction.[2] The use of the proposed automatic

[1] http://www.sail-technology.com/, http://www.quaero.org/.

[2] For instance in [4] a model of air traffic controllers conflict detection and conflict resolution that can be used in these tasks is developed, and in [5] a method to automate ATC within simulation environments is presented.

pseudopilot instead of a real person would drastically reduce the cost of such a system. Several systems are mentioned by Weinstein in [13] to justify the early military interest in this application [16–19]. More recently, other projects have worked on the same idea [20–23]. Although the controllers are expected to speak in a constrained stylized language, they will frequently stray from the constraints so it is essential for the recognition system to be able to process any deviations from the grammar effectively. In fact, according to [20], with years of practice human beings change their behavior and only a small percentage of their instructions fully conform to the International Civil Aviation Organization (ICAO) recommendations. The minimum requirement would be the need for the recognition of the deviation and request to the trainee to rephrase his speech input [13].

The system developed in this paper effectively copes with variations from the official grammar in contrast to previous systems [21, 25, 23]. If feedback from the user is allowed, one way to recognize the deviation is to use modern confidence estimation algorithms [26]. In a more recent ATC training simulator development, Adacel was using automatic speech recognition in their Adacel MaxSim training simulator. Unfortunately we do not have any further data on the evaluation of the system [27]. Several other commercial products for ATC training using speech recognition and synthesis have also recently appeared on the market such as ATVoice from UFA.[3] The brochures of the product promise appealing features such as lowering operating and recurring human resource costs, increasing efficiency and throughput during high volume exercises, and allowing the user to train independently without using extraneous resources, but we do not have evaluation data or feedback from customers to give educated advice to prospective users. The languages available are also limited: ATVoice only works for American English speech.

Other authors have proposed alternative potential applications for ATC speech recognition such as handling electronic flight progress strips [25, 28]. In a study carried out by Ragnasdottir, et al. [29], a new application for speech recognition and understanding in air traffic control (ATC) is proposed. This application is intended to support controllers in their work by making the system give warnings when a discrepancy is found in the communications between the controller and the pilot. A detailed analysis of voice communication in ATC shows that there is some sort of miscommunication in about 1% of transmissions [30]. The proposed system should recognize the read back of the pilot as a response to the ATC order and check it with its internal flight data processing system to detect errors and warn the ATC controller. The difference between this application and other applications mentioned in this section is that in this case, the speech to be recognized comes from the pilot instead of the ATC controller, thus adding problems of typical disturbances of the RF channel.

A completely different use of ATC speech recognition was made in [25] in which they worked on a project to integrate speech recognition into a C-CAST system (controller communication and situation awareness terminal) which was able to transmit, display, and receive clearances inside the aircraft through a data link channel. The objective of the system was the transcription of the speech of the ATC controller into text which would then be sent to the pilot through the data link channel.

Other potential applications include the analysis and calculation of the objective task workload of the controller by analyzing ATC speech. Some authors relate the objective task workload to both measurements of communication events and that of the variations of ATC activity (traffic complexity) [31, 32]. Communication events include time spent in the ATC-pilot communications and the content of the communications. The content of the communications can be extracted automatically by using speech understanding algorithms. It is important to note that a fully correct transcription is not required for this application inasmuch as the main keywords are detected.[4] For this kind of application a speech recognition system with a certain amount of errors can be used.

Finally, a very ambitious objective was presented in [34], in which a system was developed to embed a speech-based interface into an unmanned aircraft (UA) or unmanned aerial vehicle (UAV) that could understand ATC speech in the same manner as does a normal pilot. It controls the vehicle with the same commands used by a pilot and responds with speech synthesis with the same type of sentences that are used in a normal ATC procedure. The authors concentrated on the demonstration of only one of the en-route tasks: the flight-path-change directive. Unfortunately no quantitative data on the speech understanding performance of the system was reported.

From the aforementioned experiences, we can confirm that there is an increasing interest in learning about the capabilities of current speech understanding algorithms when processing ATC speech and exploring how to improve the ATC process.

[3] http://www.ufainc.com/brochures/ 200803%20UFA%20 ATVoice%20Brochure.pdf.

[4] An important area of research in speech recognition is that of keyword spotting [33]. By using keyword spotting systems, many understanding tasks can be carried out without the need to make a full transcription of the sentence, thus reducing the computational cost.

TABLE I
Number of Sentences used to Train Language Models for Each ATC Task, Number of Words in the Dictionaries and their Perplexity

| Task | Spanish | | | English | | |
|---|---|---|---|---|---|---|
| | Language Model Training (sentences) | Dictionary (words) | Perplexity | Language Model Training (sentences) | Dictionary (words) | Perplexity |
| Clearances | 4535 | 1001 | 15,23 | 2703 | 656 | 19,91 |
| Arrivals | 2512 | 452 | 19,57 | 721 | 267 | 16,68 |
| Takeoffs | 3717 | 753 | 11,50 | 1090 | 351 | 11,94 |
| North Ground taxiing | 12326 | 1522 | 23,19 | 2766 | 479 | 17,92 |
| South Ground taxiing | 12915 | 1612 | 29,34 | 3040 | 535 | 43,32 |

In a previous work [24] we described a summary of the results that we obtained in the INVOCA project, a project endorsed by AENA (Aeropuertos Españoles y Navegación Aérea—Spanish Airports and Air Navigation Authority) to analyze to what extent the processing of ATC speech can be done automatically with speech recognition and speech understanding systems and explore its possible applications, reporting results of the project concerning only the "clearances" task.

In this paper we describe in detail the development of the system, the five different tasks addressed (clearances, arrivals, takeoffs and the control of ground surface taxiing, divided into two areas: north and south), together with their vocabularies. All of the tasks pertain to air traffic controllers distributed between two control towers at Madrid Barajas International Airport. The methods used to carry out both speech recognition and speech understanding and, more importantly, the results of different experiments comparing training data, evaluation data, simulated task data and field data are given. Finally we discuss our results and compare them with other previously published results.

The paper is organized as follows. Section II describes the definition of the tasks of the ATC considered in the project and the data used to carry out the experiments. Section III describes technically how the system has been built. Section IV presents the experiments carried out and the results obtained. Section V analyzes the source of errors in more detail presenting a new set of experiments. Section VI contains a critical discussion and finally the paper ends with the conclusions in Section VII.

## II. DESCRIPTION OF THE TASKS AND DATA USED

The project comprised the work of air traffic controllers located at the two airport control towers in five different positions: arrivals, clearances, takeoffs, and the control of ground taxiing (divided into two different areas at Madrid airport at the time of the experiments presented here).

*Clearances*: This position authorizes flight plans, engine startup and transition to surface ground control.

*Arrivals*: This position controls the final approach phase of the plane for landing. It consists of clearance for landing, instructions on how to exit the runway, and the communication of the next control frequency for the controller: ground taxiing.

*Takeoffs*: In this position the controller supervises the takeoff process from the waiting point, entry to the runway, clearance to take off and transition to the suitable traffic control frequency.

*North and South Ground Taxiing*: These comprise the process of ground routing in the North Area or South Area of Barajas airport.

The languages to be processed were Spanish and English since the controllers at this international airport use sentences mixing both languages. The system had to detect and process both languages at the same time.

In order to develop the system we recorded many hours of speech distributed between the different ATC tasks and languages. However, most of the development, and particularly the acoustic training of the system, was carried out using only the clearances task. 7.1 hr of speech (4,026 sentences) were used to train the acoustic models of the recognizer for Spanish and 4.7 hr of speech were used for English (2,200 sentences). These sentences together with 1,531 sentences for the testing set for Spanish, 774 sentences for the testing set for English, and 1,005 sentences for the field test were the only files which our experts fully labeled due to the limited budget of the project. These full labels included the intended text (the text the expert considered the speaker was intending to say) and actual text (including labels for spontaneous language artifacts like coughs, repetitions, bad or alternate pronunciations, etc.) along with semantic interpretation labels (concepts with their attributes and values carried by the sentence). More text was obtained for language model training by simply transcribing recordings into text files (neither preparing the full labeling nor the semantic reference interpretations, and not creating the individual sentence files needed for acoustic training).

Two measurements that describe the linguistic complexity of the different tasks are given: the dictionary size of the task and its perplexity. Table I

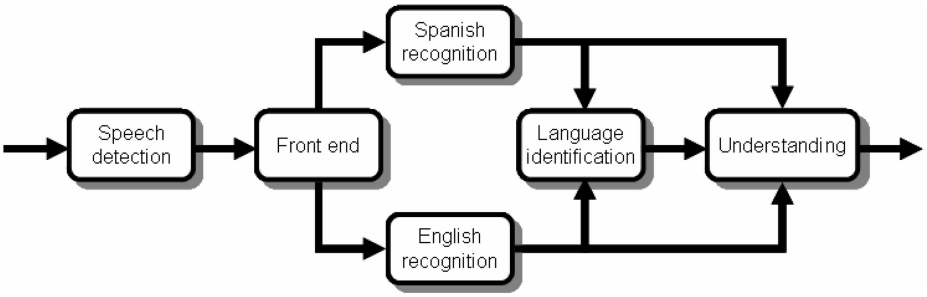



Fig. 1. Block diagram of ATC speech understanding system.

TABLE II
Dictionary Overlaps Among Tasks

| | English | Spanish |
|---|---|---|
| Accumulated dictionary | 869 words | 2086 words |
| Common to all tasks | 18.3% | 10.9% |
| Specific to just one of the tasks | 39.6% | 24.9% |

presents the number of different words for each language and task in the third and sixth column, as observed in the training material. Clearances has the biggest dictionary for English and the "South Ground taxiing" for Spanish. The lighter dictionary is found for "Arrivals" for both languages. It is known that the number of words for the ATC in the tower control is much higher than the number of words for en-route controllers, which is around 300 words [21].

Table II shows the dictionary overlaps between the different tasks. The accumulated dictionary obtained by merging the words in all tasks is 2.4 times bigger for Spanish indicating the higher proficiency and ability of the ATC controller to both use more and different words in Spanish (their mother tongue).[5] In English 18.3% of the words are common to all tasks while only 10.9% of the words are common to all tasks in Spanish. One explanation for this is the greater variability of expressions in the mother tongue. By measuring the specifity of other parts of the vocabulary, we find that 39.6% of the English words in the vocabulary and 24.9% of the Spanish only appear in one of the tasks and not in the others, so in English the controller seems to use more specific words than in Spanish for this application.

For the creation of the stochastic language models (stochastic grammars) we have used transcriptions of recorded sentences for each task. In Table I we show the number of sentences that we have used to train the language models for the different tasks together with the perplexity for each task.[6] The perplexity is a measurement of the average number of words that may follow a particular word in the language domain of the task. The perplexity is calculated from a text document. When the perplexity is low, even if the number of words in the recognition dictionary is high, the task is simpler than when the perplexity is high. Thus, perplexity is a measurement of the problem complexity. For instance the arrivals task has about half the number of words in the dictionary compared with the clearances task but its perplexity is higher for Spanish. The consequence is that the arrivals task will theoretically be more difficult to recognize than the clearances task and this fact will be confirmed in practice as we see later when we compare performances.

[5]The number of sentences used to train the systems also has some influence, but this point has not been researched.

[6]The total number of words in the training sentences corresponds roughly to 10–15 times the number of sentences.

## III. DESCRIPTION OF THE SYSTEM

In Fig. 1 a block diagram of the system is shown. In the following subsections of the content of each module is briefly explained.

### A. Speech Detection Module

This module analyzes the activity in the line and classifies it into two categories: speech and silence. It detects speech based on the energy relationship between speech and silence. Only the speech signal is delivered to the next module. This module also decides whether the pause is long enough to mean that the command has ended.

### B. Front end Processing

The speech is preprocessed to deliver a set of parameters every 10 ms. The window width is 25 ms. The parameters extracted are LPC-Cepstral (linear predictive coding-cepstral) coefficients with CMN (cepstral mean normalization) and CVN (cepstral variance normalization) [36]. As the channel has some background noise, we decided to apply these two normalization techniques, which are especially designed to compensate for channel variations. The effect of inserting a transmission channel into the input speech is to multiply the speech spectrum by the channel transfer function. In the log cepstral domain, this multiplication becomes a simple addition which can be removed by subtracting the cepstral mean from all input vectors. This is the objective of CMN: subtract the mean of all vectors. Its only drawback

TABLE III
Evaluation Results for the Off-Line Test for Spanish

| Task | Dictionary | Multiple Pronunciations | % Dictionary Words without Language Model | Perprlexity of the Test Set | Test Sentences | % Test Words without Language Model | Word Accuracy |
|---|---|---|---|---|---|---|---|
| Clearances | 1001 | 86 | 5.2% | 15.2 | 503 | 0.7% | 86.26 (±0.75) |
| Arrivals | 452 | 38 | 6.2% | 19.5 | 211 | 3.1% | 76.41 (±1.61) |
| Takeoffs | 753 | 67 | 8.9% | 11.3 | 233 | 1.7% | 85.29 (±1.25) |
| North Ground taxiing | 1522 | 86 | 5.7% | 23.9 | 349 | 0.5% | 67.93 (±1.47) |
| South Ground taxiing | 1612 | 90 | 5.6% | 29.5 | 235 | 1.2% | 72.45 (±1.5) |

TABLE IV
Evaluation Results for the Off-Line Test for English

| Task | Dictionary | Multiple Pronunciations | % Dictionary Words without Language Model | Perprlexity of the Test Set | Test Sentences | % Test Words without Language Model | Word Accuracy |
|---|---|---|---|---|---|---|---|
| Clearances | 656 | 122 | 5.5% | 23.2 | 453 | 1.2% | 73.26 (±1.11) |
| Arrivals | 267 | 52 | 3.4% | 16.7 | 57 | 0.6% | 77.45 (±3.02) |
| Takeoffs | 351 | 66 | 1.1% | 12.1 | 71 | 1.9% | 80.11 (±2.6) |
| North Ground taxiing | 479 | 86 | 1.3% | 17.9 | 70 | 0.7% | 75.90 (±3.06) |
| South Ground taxiing | 535 | 92 | 7.9% | 42.4 | 123 | 2.0% | 64.22 (±2.27) |

is that the mean has to be estimated over a limited amount of speech data, so the subtraction will not be perfect. Nevertheless, this simple technique is very effective in practice where it compensates for long-term spectral effects such as those caused by different microphones and audio channels. CVN adds a new normalization: every parameter is multiplied by the quotient of the standard deviation of the parameter in the whole database and the deviation of the parameter in the specific file. This way, the variability of the parameters throughout the database is compensated.

### C. Speech Recognition

Two speech recognizers work in parallel, one for Spanish and the other for English. We have developed a continuous speech recognizer, with HMMs (hidden Markov models) with context dependent generalized triphones with 1,500 states and 8 mixtures per state (Spanish) and 900 states, 8 mixtures per state (English) [35]. The search is driven by a stochastic bigram language model that assigns a score to each sequence of two words. These scores are learned by processing text transcribed from actual controller sentences in the development phase as we mentioned above (see Table I).

Several pruning techniques only allow our system to search through about 17% of the hypothetical full search space and respond well in real-time.[7] One pruning technique is used at the state level to avoid the computation of hypotheses that have accumulated low scores compared with the best one. The other pruning method is applied to the last state of a word with a stricter threshold. This second pruning is very relevant as it controls the number of continuation paths that will survive (and which will eventually trigger new branches in the recognition search space). The speech recognizer may use more than one pattern per word to cover several pronunciations for some words plus 14 units that we call extra-lexical units because they are models for nonlexical acoustic events (like silences, lips noise, speaker noises, hesitations like “hum,” “eh,” “mm,” etc.) that do not follow grammar rules in their occurrence probability [35]. In the third column of Tables III and Table IV

[7]0.63 times real time for the longest Spanish clearances task on an AMD Athlon (tm) XP 1800+ with 1.5 G RAM.

the number of word models to cope with multiple pronunciations is presented for Spanish and English. The 4th column of Table III and Table IV shows the percentage of dictionary words that did not appear in the training text so there is no stochastic language model for them (although their pronunciation is included in the dictionary of available words). They are given an intermediate score (the average between the largest and the shortest values in the language model) when they form part of a sentence.

The output of the Spanish and the English recognizers is a set of words corresponding to the best hypothesis that the system attributes to the pronounced sentence together with an acoustic and linguistic model combined log-likelihood score for the whole sentence.

### D. Language Identification

To carry out language identification we considered several alternatives. We have to take into account that the characteristics of this task make it particularly difficult as the controllers are nonnative English speakers. Moreover, the domain vocabulary includes words which do not provide clear evidence to distinguish which language they were pronounced in, like: alpha, bravo, charlie, some city names, airline names, types of aircraft and others with a very similar pronunciation for both languages. Furthermore, controllers often mix both languages in the same sentence, most of the times for greetings, for instance saying "buenos días" (good morning) in Spanish while the rest of the phrase is pronounced in English.

Our final choice was to base the identification on the score given by the full continuous speech recognizer for both languages running in parallel. As we demonstrated in [37], the results obtained with this technique are probably the best that can be obtained, as it models both acoustic and phonetic information, together with the sequence of allophones and words. However there are several disadvantages: a complete speech recognition system has to be trained, a lot of labeled data is needed and it would be difficult to have a real-time system for several languages as the full recognizer is more time consuming. In any case, for the identification of two languages, as in our case, it is the best option with a low error rate for both languages and it is extremely important to obtain a very good rate because errors in language identification cannot be corrected later in the system. In [38] a full recognizer is also proposed and the recognizer scores are normalized and compared with a linear classifier.

Another typical approach seen in the literature is the so-called "phonotactic approach," which classifies languages based on the statistical characteristics of the allophone sequences [39]. The technique is called PPRLM (parallel phone recognition language modeling) and its main objective is to model the frequency of occurrence of different allophone sequences in each language. The system has two stages: in the first stage, a phone recognizer takes the speech utterance and outputs the sequence of allophones corresponding to it, this sequence is then used as input to a language model module; in the second stage, the language model module scores the probability that the sequence of allophones corresponds to the language. The performance of PPRLM is lower than the method of using full recognizers.

We could also have tested other approaches based on acoustic features which are derived from the speech signal itself, such as mel-frequency cepstral coefficients (MFCC) or shifted delta cepstral (SDC) features produced by applying a 7-1-3-7 SDC scheme [40]. The reason is that the distribution of acoustic features reflects the statistics of the sound distributions in a particular language. They have been applied using modeling techniques such as Gaussian mixture models (GMMs) [40] and support vector machines (SVMs) [41]. Although acoustic features can be easily obtained from the speech signal, the useful language information is often corrupted by the distortion caused by the transmission channel or speakers. So, many studies have focused on improving the expressiveness of acoustic features for language characterization and to compensate noise and distortion [40, 42]. The results are comparable to the PPRLM technique, as we can see in [40]: the system using GMM plus SDC features obtained worse results than PPRLM, and only the fusion of both systems provided small improvements.

### E. Understanding

The understanding module processes the output words of the recognizer and obtains its conceptual content, taking into account the key concepts of the task. The algorithm builds the meaning in an island-driven bottom-up approach making use of context-dependent rules. It differs from more traditional approaches in two main points: first, we do not use formal grammars like the recursive transition networks of Carnegie Mellon University's Phoenix [43] used in several successful applications or finite state machines of the system used by Duke, et al. for the Pinocchio UA control system [34] that both need the expansion of the concepts into word constituents. In our case, it is the conceptual tagging of the words processed by the proper set of rules that elaborate the meaning of the sentences. These rules are similar to a parsing grammar, but the power of our system is the possibility of using rules that are dependent on the context and that this context can be expressed in various ways including far-reaching context. The other point different from the traditional

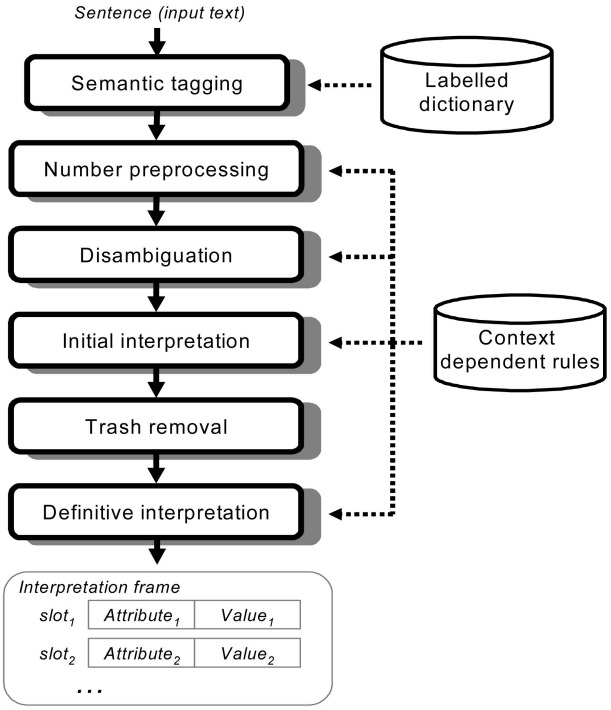



Fig. 2. Block diagram of understanding modules.

understanding methods is the use of ambiguity in the semantic tagging (each word can be associated with several tags) and the use of the "trash" tag as another tag possibility as further explained below. Like other laboratories (including [34] and [43]), our understanding module does not try to understand every single word in the sentence but tries to extract as many islands of correctly understood segments as possible. A task-specific and language-specific dictionary is needed, while the interpretation rules for each task are quite language independent. In very few cases we have used language-dependent knowledge in the elaboration of the sequence of understanding rules because our design operates on a very high (conceptual) level of information.

In Fig. 2 we show our understanding architecture. The process for each sentence begins with the labelling of each word using a set of semantic-pragmatic categories (semantic tagging). Several tags can be associated with the same word. The selection of the tags is task dependent. All the words that do not provide information are tagged as trash. The system also assigns an on-line trash tag to the new words found in the recognition evaluation experiments. This feature is also kept for final evaluation sessions of the whole system.

The algorithm proceeds by processing sequences of numbers and translating them into figures. This step has special characteristics, dependent on the specific phraseology used. The following step tries to minimize the level of ambiguity in the tagging of the words. Context-dependent rules are especially suitable for this task because the reason for selecting some tags and rejecting others for a particular word is found by looking at the presence or absence of other tags in the sentence that matches or contradicts a particular interpretation or function of the word. Following the disambiguation step, we generate the initial interpretation. Making use of context-dependent rules again we try to form islands of interpretation. These islands are often tagged with brand-new tags that do not exist in the original labelled dictionary. These new tags help the system recognize the formation of these islands of interpretation (reliably understood parts). Interpretation islands can be combined together or with single words to build larger islands of interpretation. When no more work can be done, we remove all words that are uniquely labelled as trash. The reason for removing them at this time and not earlier is that although we cannot extract any meaning from a "trash word," it sometimes helps us to define frontiers between blocks whose constituents should be jointly interpreted but separately from others. This is only true for the kind of rule that works with "near context" and that do not cross these borders. We also use another kind of rule that looks for context anywhere in the sentence.

After "trash removal" we run the definitive interpretation stage in which another group of context-dependent rules carry out their work. This part of the interpretation module is written taking into account that the trash words have being removed and special care must be taken in order not to mix things up that were previously separated by trash. The final product is a frame containing a variable number of slots, each made up of an attribute and a value that represents the interpretation of the sentence.

In Fig. 3 we show an example of a context-dependent rule and its application in the context of the ATC clearances task. We have a set of 12 "primitive rules or functions" that make up a specific language in which we write our understanding modules.

The clearances task (the most complex conceptually) is dealt with by using an initial interpretation module made up of 56 rules and a definitive interpretation module, after trash removal, with another 32 rules. In the example detailed in Fig. 3 we show a rule that relies on the consecutive appearance of two segments labelled LABEL1 and LABEL2. If this condition is satisfied, the rule selects just one of the two items as specified in the N parameter (the other item is removed from the working space) and labels it with the label specified in NEW_LABEL. Thus, this rule (function) has four parameters that have to be specified. For this rule the context is a near context (indeed a consecutive context), but we use also rules able to analyze far contexts. In Fig. 3 we show two applications of this rule.

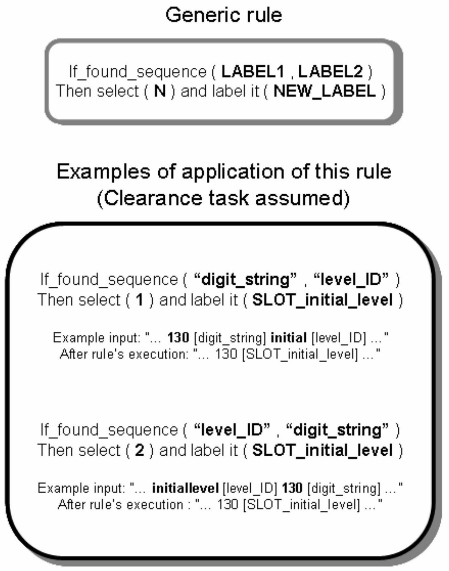



Fig. 3. Understanding rule example.

In the first example, if the sentence elements present at the moment of the execution of the rule include "…130 initial…," the rule selects the first element ("130") and labels it as "SLOT_initial_level." Something quite similar occurs for the second example. The difference is that in this case a change in the ordering of the elements produces the same result. The element "initiallevel" is obtained by previous rules that group the two words together into just one merged element labelled as level_ID.

One advantage of using context-dependent rules instead of finite state machines (FSMs) such as the ones used in [34] is, that with a proper sequence of rules, the system is able to understand some expression variants (that although not canonical with respect to the official phraseology, they do occur in real spoken examples since ATC speech strays from the canonical model). Another property of our system is its robustness against recognition errors as long as the system tries to solve all the islands of possible interpretations and does so by relying only on content words. The designer of the understanding module has to bear this in mind and should not use rules dependent on words with a high probability of generating a recognition error. These are, for example, short words without a crucial meaning in the application considered (like articles or other function words).

The design of the understanding module is quite easy for experts in the domain once they get a feeling for the set of rules (or functions) available (the "primitive" functions) and they follow some guidelines that we have learned from the experience obtained after applying this procedure to different domains. One guideline is to think thoroughly about the tags for the words. The system allows the processing of multiple tags for each word and this is relevant when designing the labelled dictionary mentioned in Fig. 2. The different tags that the designer writes down for each word should consider the different meanings of the word in the particular domain. One interesting possibility for a word is to express different tags corresponding to different meanings including the "trash" tag for words that may or may not have a meaning in a particular sentence. The context-dependent rules, mainly those present in the disambiguation module, will try to refine the multiple tagging by selecting the one (or the ones) most suitable for a particular sentence. Another key design guideline is to apply specific rules before general ones. This is necessary because specific rules try to match a context with more conditions. If a general rule is applied beforehand, it will be used because its general condition will also be met, often causing the later specific rule to be unable to find its specific context conditions. This will lead to a more general, less precise interpretation of the sentence, leaving elements out of the interpretation and eventually producing conceptual errors.

Fig. 4 presents an example of the output of the speech understanding module. In the first part of the figure, the result of the speech recognition system is shown. The set of words delivered by the recognizer is processed by the understanding module providing a set of slots with attribute-value pairs.

## IV. EVALUATION EXPERIMENTS

### A. Off-Line Evaluation

The first tests carried out on the system were the off-line evaluation tests. This is the kind of evaluation

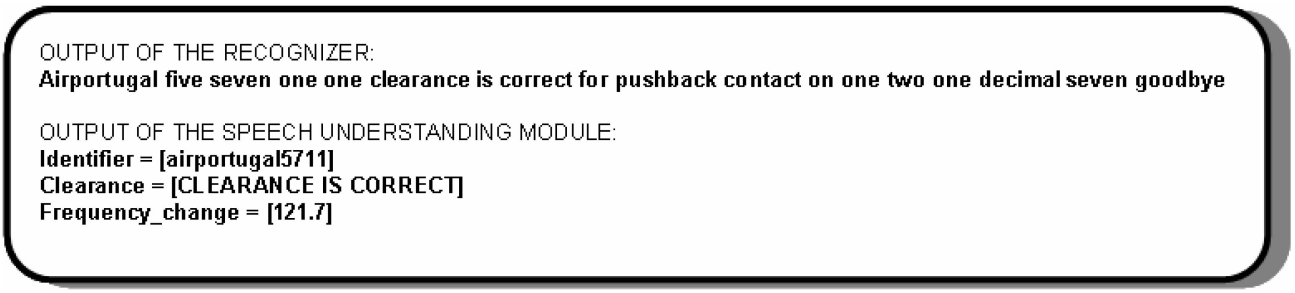



Fig. 4. Example of output of speech understanding module.

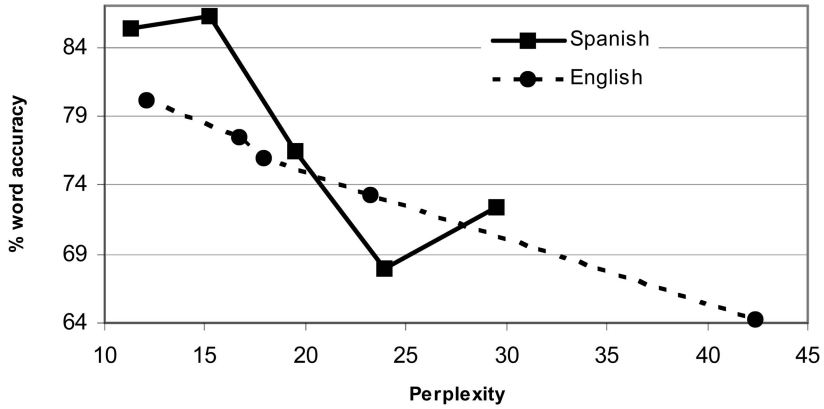



Fig. 5. Word accuracy versus perplexity for Spanish sentences and English sentences.

that the experts apply regularly in the laboratory to predict the performance of any system. For this test, a set of sentences that were not used in training the recognizer, neither for the acoustic models nor for the language models, was fed to the system and the results were computed. The effect of the language identification algorithm was eliminated in these tests and will be evaluated in a different test.

1) *Speech Recognition*: A usual measure of the performance of a system is given by the word accuracy rate that accounts for all types of errors (substitutions, insertions and deletions) compared with reference data. The formula used to calculate the error rate is as follows:

$$\%\ \text{word\_error\_rate} = \frac{\#\ \text{substitutions} + \#\ \text{insertions} + \#\ \text{deletions}}{\#\ \text{referencewords}}$$

$$\%\ \text{word\_accuracy} = 1 - \%\ \text{word\_error\_rate}. \quad (1)$$

The number of substitutions, insertions and deletions is calculated with a program that finds the best alignment between the hypothesis sentence and the reference sentence by considering a unity cost for each of the three kinds of error. The last column in Table III and Table IV gives the word accuracy of the system for the different tasks and for Spanish and English. Data is given with 95% confidence intervals (in parentheses), as in expression (2).

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} \quad (2)$$

where $p$ = % word accuracy and $n$ = the number of reference words.

These tables also show the perplexity of the test set, the number of test sentences and the percentage of test words that do not have a language model because they did not appear in the training text. It can be seen that there are significant differences in performance in the different tasks. These differences are mainly due to the perplexity of the tasks.

In Fig. 5 word accuracy performance is plotted against perplexities for different tasks. It can be seen that in general the performance decreases as the perplexities increase.

In Spanish, slight variations in this general trend are obtained for perplexity 15.2 and perplexity 29.5 which are higher than the trend. One factor for this improvement is that acoustic models were only created with sentences that come from the clearances task which is the task with a 15.2 perplexity value and the percentage of test words without language model is small. In the second case (perplexity 29.5, South Ground taxiing) one reason for the improvement is that the language model in this case is better trained because it has the greatest number of sentences to train it (12,915 sentences).

A source of high perplexity is either a very variable grammar or a not so variable grammar but in which some elements have high variability. This second example is the case for taxiing tasks in which the number of different expressions is not as high as in a clearances task (for instance), but where the number of different paths, specified as a sequence of fixed points and routes, is high. For English, the effect of perplexity is also clearly observed. A conclusion is that, as was expected, for ATC speech the performance of the speech recognition system depends on the perplexity of the task, thus, its perplexity is a good prediction variable.

Fig. 6 plots a comparison of results for Spanish and English task by task and for all tasks and the weighted average which is calculated giving a weight proportional to the number of test sentences for each language. The plot also shows confidence margins for each data.

The comparison between English and Spanish for each task shows, in general, a significant better performance for Spanish. There are multiple causes to justify this fact: we had more Spanish data recorded both for training acoustic models and for training language models; we have more experience building Spanish systems and more knowledge of the language,

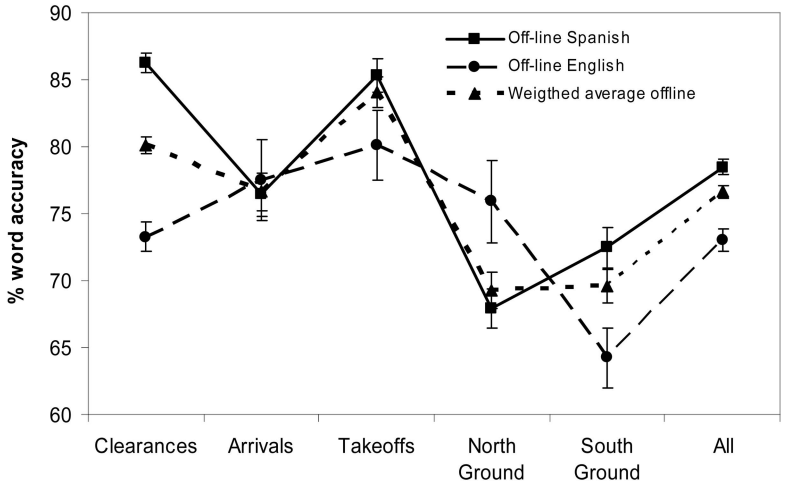



Fig. 6. Offline word accuracy results for different tasks with all of them together and in both languages. Weighted average is calculated giving weight proportional to number of test sentences for each language.

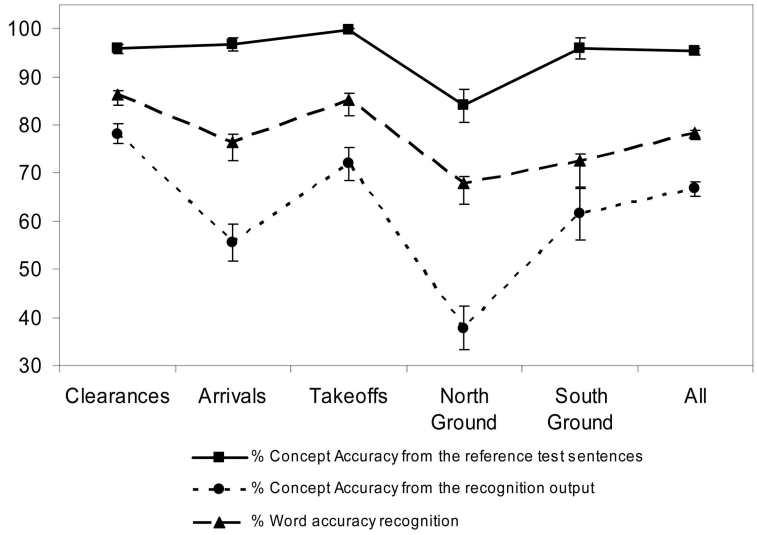



Fig. 7. Offline speech understanding results for Spanish, comparing results using reference words and using output of recognition module. Middle curve reproduces word accuracy after recognition module.

which influences our ability to describe optimal phone inventories, multiple pronunciations, etc.; and last but not least, English examples are uttered by nonnative speakers with a very high pronunciation variability. One reason for the differences from Spanish to English for the North Ground taxiing task (better performance for English) is the fewer number of words in the dictionary for the English task and the lower perplexity of the language model compared with Spanish.

2) *Speech Understanding*: Table V presents the number of semantic slots evaluated for Spanish and English. The number of slots is important to establish the confidence intervals of the results. Fig. 7 presents the percent concept accuracy, across different tasks for Spanish. Percent concept accuracy is calculated in a similar way as percent word accuracy calculated in the previous section. The only difference is that when we calculate the match between our system output and the reference, we require the coincidence of both the attribute and the value for each concept to count it as a positive match. The top line represents the results of the understanding module on the reference sentences (the transcription of the test sentences made by hand). This line shows the power of the understanding rules by themselves which, in general, cover the domain of the application quite well (with

TABLE V
Speech Understanding Slots Evaluated for Spanish and English

| Task | Slots Evaluated (Spanish) | Slots Evaluated (English) |
|---|---|---|
| Clearances | 1545 | 1032 |
| Arrivals | 621 | 165 |
| Takeoffs | 655 | 207 |
| North Ground taxiing | 439 | 86 |
| South Ground taxiing | 297 | 149 |

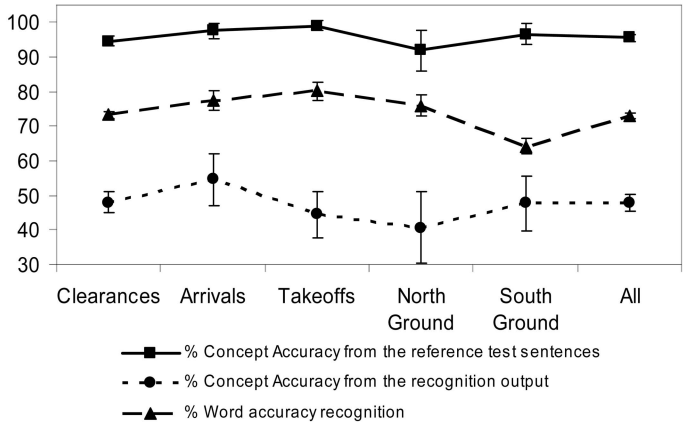



Fig. 8. Offline speech understanding results across tasks for English, comparing results using reference words and using output of recognition module. Middle curve reproduces word accuracy after recognition module.

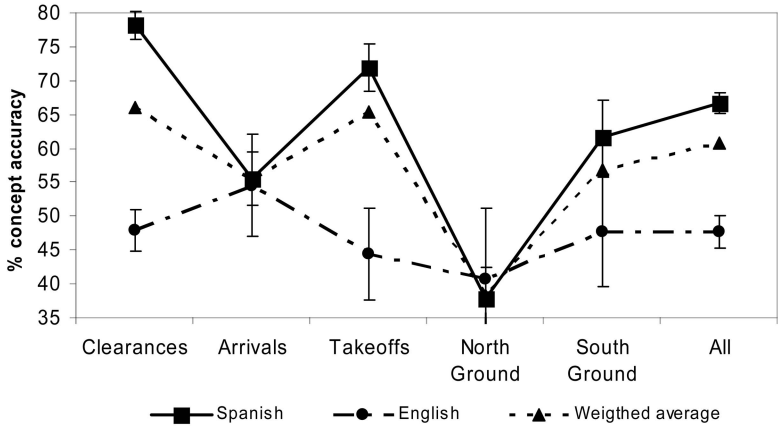



Fig. 9. Comparison of off line speech understanding results for Spanish and English for different tasks and all of them together. Weighted average is dependent on number of slots evaluated.

a slightly lower performance for the North Ground taxiing task, 84.05%, due to a different sharing distribution of the errors across the dictionary. In fact, for these cases, word accuracy is a misleading predictor of understanding performance, as the accuracy on the "carrier phrase" words is less relevant than the accuracy on the particular paths mentioned in each case, which have to be correctly understood.). The average performance of the understanding module is 95%. The lowest line in Fig. 7 shows the understanding performance after the recognition stage. The middle line in the figure plots the word accuracy from the recognition stage. We notice that the recognition errors made by the recognition module have an effect on the final results, thus increasing the concept error rate. The general performance for each task follows the performance trend obtained from the speech recognition results, showing the expected correlation between recognition (text transcription) and understanding (semantic content extraction) capabilities. In Fig. 8 the concept accuracy from the text (upper line) from speech (lower line) together with the word accuracy (center line) are plotted for English. The recognition errors coming from the recognition stage are augmented in the understanding stage as it is in Spanish. However, as we see later in Fig. 10, if we consider the number of correct sentences the understanding module improves the performance of the recognition module.

In Fig. 9, a comparison of speech understanding results for English and Spanish is given across tasks and for all the tasks together with their weighted average. The general trend observed in speech recognition results is also observed here. Spanish delivers better results in general than English although in two cases (Arrivals and North Ground taxiing) there are no statistical differences. Notice that the speech recognition results for Arrivals were also similar for Spanish and English (see Fig. 6).

Another way of analyzing the results of the speech understanding system is by calculating the number of sentences with no errors (perfect sentences). While in speech recognition, a single word error in a sentence causes a sentence error. In speech understanding a single word error may not cause a concept error. This happens because some of the words that might be erroneous in the recognition result are not used for understanding (i.e., a function word like an

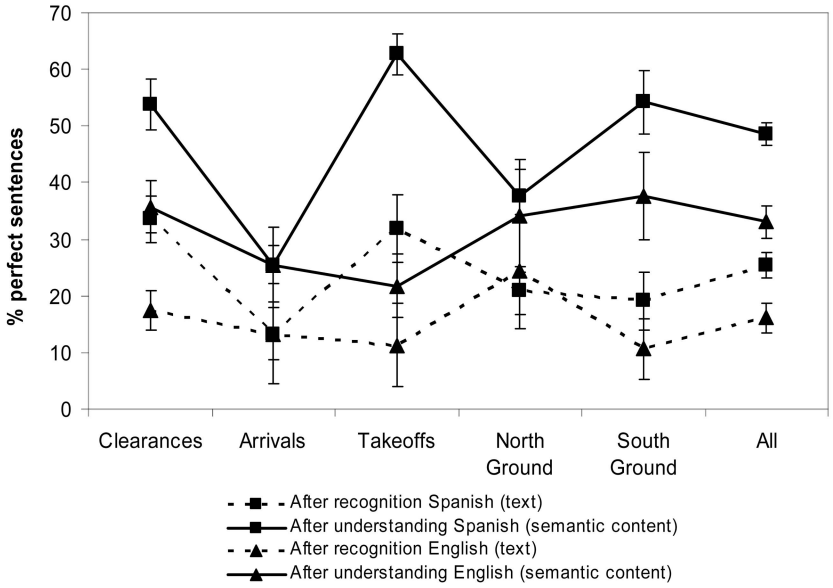



Fig. 10. Comparison between percentage of correctly recognized sentences and correctly understood sentences for Spanish and English across tasks and all of them together.

article or preposition) and other crucial words give the necessary conceptual imprint to form a correct interpretation. Thus, the speech understanding module "corrects" the output of the speech recognition module to a certain extent. This is an important robustness characteristic of our understanding solution. In Fig. 10 the percentage of correctly recognized sentences for English and Spanish is presented together with the percentage of correctly understood sentences across tasks and all together. All the tasks exhibit the same trend: the percentage of correct sentences is improved. For Spanish this improvement is on average 25% in absolute points and for English it is a little lower (17% on average) but it is also true that the base performance of the English recognizer is lower than the Spanish one and a sentence full of recognition errors is very difficult to "correct" by the understanding module. Depending on the use of the system, the average 50% fully correctly understood sentences may be enough for some applications (for example, if one wants to detect the workload of the controller).[8]

3) *Language Identification Module*: The language identification module was evaluated offline using a small set of 60 sentences. The percentage of correct sentences was 96.67% and 3.33% of error.

## B. Field Evaluation

After testing the system in the laboratory with original recordings (field recordings) made at the beginning of the project we carried out a live test (i.e., a test carried out with the system connected to the microphone of ATCs on duty). This test is the one that gives the real performance of the system. With this test, we check the performance of all the modules working together in real time, the Spanish recognizer, the English recognizer, the language identifier, and the understanding modules. In this evaluation both languages are mixed. The results that we present are the overall results for each of the tasks. Table VI summarizes these results for the speech recognition part. In the live test, new words appear that are not included in the dictionary of the system, these words are called "out of vocabulary words" (OOVs) and are presented in Table VI in the third column (the OOVs as a percentage of words in the dictionary) and fourth column (the OOVs as a percentage of words in the test). New words cannot be understood by the system because they are not known to the speech recognition

TABLE VI
Evaluation Results for the Field Evaluation Test for Live Conversations (Languages Mixed)

| Task | # of Sentences | % OOV in Dictionary | % OOV in Test Words | Word Accuracy |
|---|---|---|---|---|
| Clearances | 385 | 2.72% | 1.15% | 77.99 (±1.05) |
| Arrivals | 158 | 1.95% | 1.07% | 79.85 (±1.63) |
| Takeoffs | 167 | 2.17% | 1.29% | 76.16 (±1.79) |
| North Ground taxiing | 206 | 0.9% | 0.99% | 66.24 (±1.89) |
| South Ground taxiing | 89 | 0.14% | 0.67% | 62.95 (±2.94) |

[8]The only way to know if it is enough for the application is to build the application and carry out usability tests.

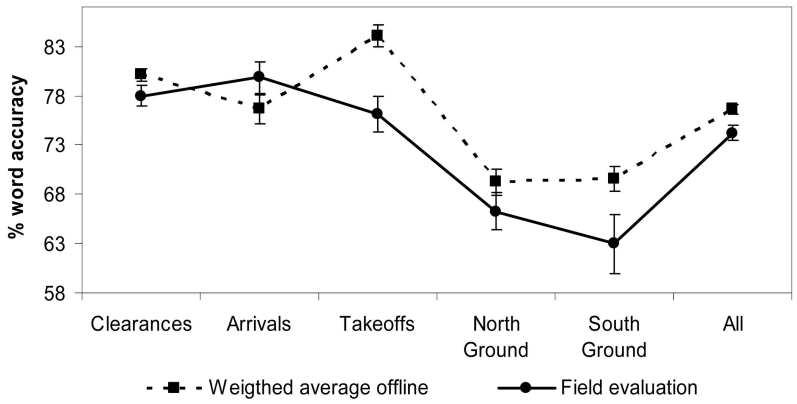



Fig. 11. Comparison between weighted average results and field evaluation results.

module nor to the understanding module in advance and they reduce the performance. The percentage of OOV words is never zero even for these tasks in which the phraseology of the communication has been designed to be standard. On the one hand, human communication eventually produces a relaxation of the norm that alters the expected grammar and, on the other hand, there are other parts of the message (the variable part dependent on the particularities of each airport) that are not specified in the regulation nor are they kept constant because of the natural dynamic changes in time (new runways with their new taxiing routes, their new identifiers, etc.). Even after several thousand training sentences, we find OOV in test sentences.

In Fig. 11 we compare the results obtained in this section with the results obtained in the weighted average evaluation made in previous sections that we have called offline. The use of the weighted average is needed since in the field evaluation both languages are mixed. If we analyze the results presented in Fig. 11 we can see that field results are lower on average than the offline results in four of the tasks. The lower performance for the 4 cases is on average 5% which is influenced by errors due to the language identification module (see Table VII), OOVs, and the speech end point detector. If the language identification module has an error, the full sentence is wrong since all the words recognized are wrong (they are in a different language).

To further analyze the performance of the system, we calculated the percentage of errors of the language identification module in the live test. The results are presented in Table VII. Although the errors for English identification are greater than those for Spanish identification, the weighted average performance is 95% (or 5% error rate), not far from 3.3% in offline tests.

The performance of the speech recognizers both in Spanish and English follow the expectations obtained in the offline evaluations. The reason for a higher performance for the "arrivals" task is still to be researched. On average, the results of the field test are only lower than the offline tests in 3.3%, but if we exclude language identification errors the field tests outperform the offline tests in 1.7% absolute points due mainly to the results of the arrivals task. This result is not strange, since the offline data are also field data recordings. The main difference is that offline data is processed in the laboratory and the field data is processed in real time, with the computer connected to the microphone of the controller.

TABLE VII
Performance of the Language Identification Module in the Field

| Identified Language (# sentences) | Language Spoken (sentences) | | |
|---|---|---|---|
| | Spanish | English | Weighted Average |
| Spanish | 757 | 36 | |
| English | 17 | 253 | |
| Percent correct | 97.8% | 87.54% | 95.0% |

In Fig. 12, the understanding results in the field are compared with the weighted offline understanding results. Again the performance of the field results is lower than the performance of the offline tests for three of the tasks and not significantly different for two other tasks. In this case, the relatively low performance of the speech understanding for the North Ground taxiing task, coming from the relatively low Spanish speech recognition results for this task and the relatively lower performance of the understanding rules is not degraded any more in the field test, possibly indicating that the errors in the language identification have their origin mostly in the takeoffs task and the South Ground taxiing task. The percentage of OOVs here also has an important impact in the degrading of the results particularly for clearances and takeoffs. A single OOV causes a concept error while in the offline tests there were no OOVs in the understanding dictionary.

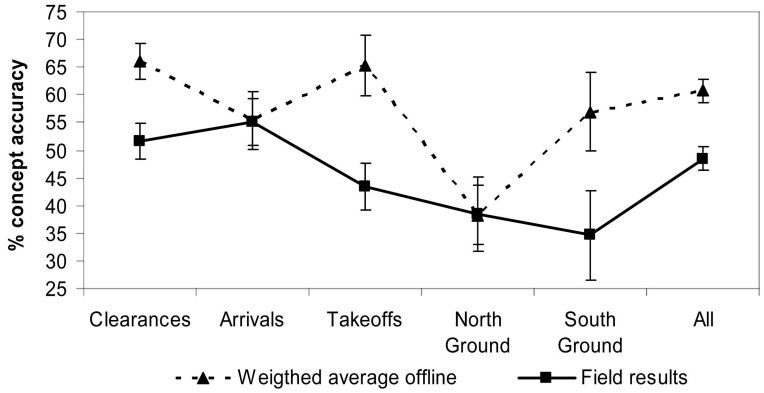



Fig. 12. Comparison between weighted offline understanding results and field evaluation understanding results.

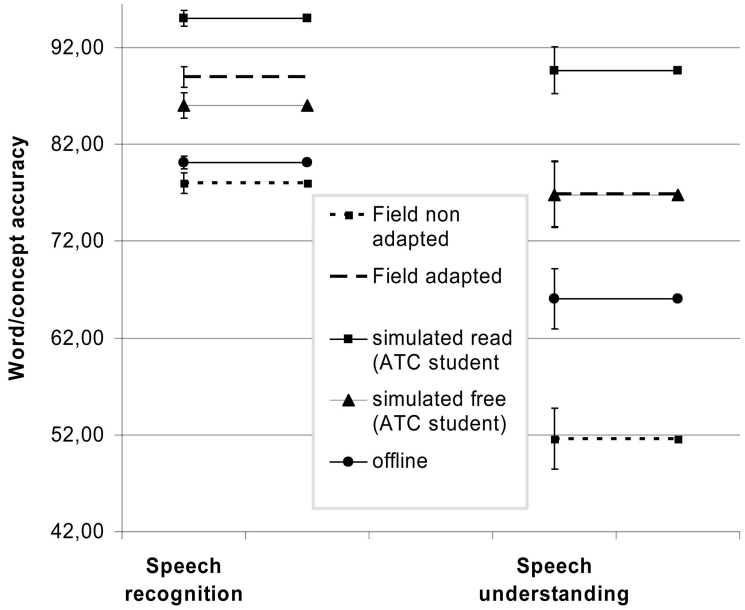



Fig. 13. Results for the same task (clearances) and changing conditions.

## V. PERFORMANCE ANALYSIS

We have further looked into the possible source of errors by analyzing the clearances task [24]. A detailed analysis of the training data and the test data showed that there are two situations at Barajas airport. One appears when the wind comes from the north (called the north configuration) and the other is when the wind comes from the south (called the south configuration). While our original recordings all contained data for the north configuration (the most common one), the field experiments were carried out by chance with the south configuration. This is a reason for a degraded performance in the field tests. We made new recordings using the north configuration and the speech recognition results improved by up to 88.96% word accuracy and 76.87% concept accuracy (see Fig. 13, "field adapted" results compared with "field nonadapted" both for recognition and understanding). Fig. 13 also shows offline results for completeness. The conclusion is that the data capture is especially important for ATC as a change in configuration of the airport or the loss of a particular circumstance will cause relevant degradations in the system's capabilities since the stochastic grammars that we use need training sentences. The help from ATC experts is needed in order to maximize the suitability of the data capture and repeat the capture several times in different conditions to ensure a reasonable sampling of the variability in a real system development.

The next experiment was carried out in a simulated task experiment in which 7 ATC students freely generated a set of sentences based on a given operating scenario. The experiments were carried out separately for English and Spanish so no language identification was used. The weighted results are presented in Fig. 13 (called "simulated free (ATC student)"). For speech recognition the results are significantly lower than the field adapted case. A more detailed analysis comparing results for Spanish and English showed that the performance of Spanish recognition was not significantly different from

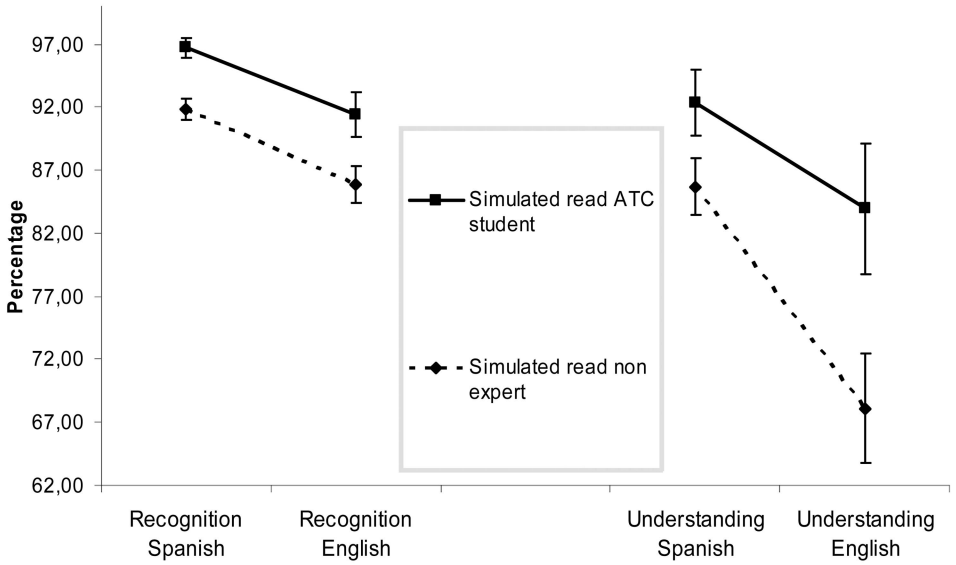



Fig. 14. Results for clearances task and simulated (read) speech with ATC students and nonexperts.

the field adapted case while for English the results were significantly lower and this fact weighted the combined results downwards.

However, for speech understanding there are no significant differences between simulated free experiments and field adapted experiments. A closer look at the English and Spanish results showed an improvement in the Spanish results compared with the field adapted case but a significant lower performance for English that compensated both effects. The understanding results of the simulated free experiments for Spanish demonstrate that a less spontaneous pronunciation (such as was heard in the experiments) improves the understanding results (81.77% concept accuracy versus 76.87%, although not significantly). One reason for the lower performance of English both in recognition and understanding is due to the lower English language skills of the ATC students. We should also remember that the field adapted case experiment includes the language identification module so the comparison between these results and simulated task results is not direct.

The results of the simulated free experiments constitute the best prediction of what one could expect in an ATC training simulator application since there is a free framework in which the ATC student has to perform a defined task.

A third experiment was carried out by giving the speakers several sentences to read (simulated read ATC student in Fig. 13). The performance both for speech recognition and understanding is better than the field adapted case, however, read speech has a much lesser spontaneous style. Our conclusion is that spontaneity in the live speech is difficult to understand compared with speech obtained in a "more controlled" experiment (i.e., simulated read) as is well known in other speech applications. The results obtained from read speech can be used as an indicator of maximum performance expected under the best circumstances.

Finally the same sentences given to the ATC students were read by 16 speakers not familiar with ATC phraseology chosen from among people working in our laboratory. The reason for carrying out this experiment is to highlight that when designing a system for ATC, it is crucial to work with ATC professionals and field data and not to rely on informal tests with nonprofessionals. This concept is very well known in the speech community but we have the experience that it is not observed in other disciplines in which speech technology is applied without taking these design details into account.

The results for speech recognition and understanding for both languages are presented in Fig. 14 called simulated read nonexpert compared with the simulated read (ATC student) of the previous paragraph. The results for the more experienced users (ATC students) yield significantly better results both for English and Spanish. This is due to the different style of pronunciation obtained from people who are not familiar with the task (nonexperts) compared with the style of ATC controllers present in the training material. The ATC students present a style closer to the professional speech in the training material.

The experiment highlights two issues: that professional speech is quite different from nonprofessional speech and that our design procedure ends up obtaining a system well adapted to professional speech, giving worse performance for nonexperts inasmuch as they are unable to reproduce all the characteristics of the professional speech.

In conclusion we confirm for the ATC domain that in order to determine the performance of the system, the best approach is to use experiments from the field

and that special care has to be taken in order to cover all different conditions when capturing the training and testing data. A good estimation of performance can be made by creating a simulated scenario in which the user has to perform a task and is free to use whatever sentences he wants. Experiments using read sentences demonstrate the maximum capabilities of the recognizer or understanding module in the domain but are not a good estimator of the final expected performance.

## VI. DISCUSSION

Commercial off-the-shelf (COTS) systems as used in previous experiences [20, 44] need a specific grammar that has to be developed with a great deal of effort and is never complete. Out-of-grammar sentences lead to big errors. With a design customized to the task, as is done in this work, results can be better and more robust if automatic learning techniques are used. For these stochastic schemes, the quantity and completeness of data available for training is relevant to determine the resulting performance.

Our speech understanding architecture, based on a bottom-up island-driven approach using context-dependent rules, exhibits robustness against recognition errors and at the same time easily accommodates different orderings for the sentences like those present in real ATC conversations. Its design is easy for the experts once they follow a few sensible guidelines in the process.

If we compare results for different tasks, they are diverse as a result of several factors, the main one being the different perplexities of the tasks. The different number of sentences used to train each task (for both language and acoustic models) can also be a source of lower performance for some tasks although this has not been researched. Results obtained by comparing offline data with online data are not significantly different when taking into account the errors made by the language identification module and the set of new words that appear in the experiments. From the analysis of fully correct recognized sentences and fully correct understood sentences it is important to point out that while the speech recognition module makes errors, some of them can be corrected by the understanding module. Results comparing online data with simulated task data (read) demonstrate that the style used by ATC controllers in live speech is much more difficult to recognize than the same sentences read in a controlled simulated task experiment. The familiarity of the user with the task is also a positive factor in terms of recognition performance compared with users that are not familiar with the tasks and sentences.

It is difficult to compare our results with previously published results since the conditions are different. For instance in [21], in the context of an ATC training simulator development in English and French, it is reported that 95% of the sentences were correctly recognized and understood but it was done with a task covering only 250 words and a finite state grammar. The results presumably were also laboratory results since they mention that evaluations by ATC people were on their way. The problem with finite state grammars is that the contents of the grammar cannot vary online in the final system, which implies that grammatical variants that were not considered during the development of the grammar will not be allowed by the recognizer. Instead, if stochastic grammars are used, although a particular grammatical variant was not considered (observed in the training material), the recognizer is nevertheless able to recognize the new variant because the stochastic grammar will give a score (although most of the times lower than when the utterance grammar matches the observed data) that is not zero (thanks to the smoothing techniques). Thus the recognizer gives the right answer if the acoustic evidence is enough. In this sense, the stochastic grammar exhibits a robustness behavior allowing unforeseen grammatical structures not allowed with finite state grammars. This characteristic adds up to the fact that although we are dealing with professional speech that should keep specific grammatical constrains, humans are unable to strictly keep to this normative language and always stray from the official phraseology. The result is that this robustness characteristic of our system is crucial for speech technology use in ATC. A similar conclusion holds for the specific understanding technology that we decided to use for this work as far as it is powerful enough to admit changes in the ordering of appearance of the elements that build up the final meaning of each sentence and is even able to jump on some irrelevant words when building the interpretation islands.

A closer comparison with our experiments was carried out by Hering [20] because he carried out experiments with live recordings during a simulation exercise including hesitations, insertions in a different language and recordings that do not strictly respect the phraseology. He tested three COTS systems and reported the value of the percentage correct of between 26% and 39% in the recognition task—there was no understanding module—and he used a different way of measuring the error rate (he did not count insertions). The tested task was an en-route task with a 300-word vocabulary which had 11.2% of OOV words. These results were obtained from a simulation exercise and not from live recordings.

There is still room to improve the current performance of the presented ATC speech understanding system. We could get a significant improvement just by applying pragmatic constraints as is done in [20]. These data could be incorporated

into our system as a set of restrictions which, in short, would mean a lower recognition uncertainty and therefore better recognition and understanding accuracy. We are referring to, for example, the knowledge on the set of available communication frequencies and runways, the list of possible call-signs, flight levels, etc., constraints that have not been applied to our system (i.e., all possible combinations of numbers are recognized, but not all numbers are possible when speaking about frequency changes).

The application of speech recognition and understanding methods to ATC speech has shown a varied range of results across tasks and languages. The use of these algorithms in a real environment depends on the requirements of the application. While a very demanding application (i.e., fully automating the process of ATC) requires a better performance, many other practical applications do not, particularly if a confirmation mechanism is used. Although it has not been implemented here, the confirmation mechanism—usually a confirmation question asked by the system—is generated automatically by the system when it perceives that the confidence of the results is low and the user has to repeat the command (this is a credible situation in the context of ATC-student training). In this last case, even a certain level of error is useful in order to better simulate an understanding problem with a pilot or with the communication channel. Another example of a possible application is scoring ATC student speech for his or her training. Finally an ATC task workload analysis, ATC controller's performance measurement, or detecting possible miscommunication errors between the controller and the pilot are several feasible applications using today's state of the art systems.

The only way to be sure about the level of speech recognition and understanding performance needed for a particular application is to implement it and involve users giving feedback on the usability of the product. But again this is very dependent on how the application is built and not only on the recognition or understanding results. This objective was out of the scope of our project.

The fact that there are products for training ATC using speech recognition (with no or very limited understanding) lead us to conclude that the performance of current systems is enough for this kind of application. But the question "is it enough?" has to be asked to the users. Unfortunately no data is published on the usability of these applications.

## VII. CONCLUSION

In this paper we have drawn up a revision of experiments and experiences carried out in the literature to process ATC speech automatically.

We have also described in detail the results of the INVOCA project, a project whose objective was to analyze to what extent the processing of ATC speech can be done automatically with current speech recognition and speech understanding systems. The system is able to process sentences whose content is mixed in Spanish and English. We have presented our methodology used to do both. In comparison to previous systems, by using stochastic grammars our speech recognition algorithms allow the processing of sentences that stray from the ICAO standards as is often observed in real ATC speech and does not need a lot of effort in writing always incomplete specific grammars. However, the recording of data from the field is needed. The analysis of captured vocabularies renders 2,086 different words for Spanish compared with 869 different words in English indicating a higher proficiency of the ATC controllers in Spanish. We have also developed a robust algorithm for speech understanding that allows the flexibility of working with sentences with no restrictions on concept ordering. The average performance of the understanding algorithm from text sentences is 95% both for English and Spanish demonstrating that the understanding module is robust. We have also demonstrated that the understanding module is capable of improving the recognition performance when we compare the number of fully correct sentences obtained after the recognition module to the number of fully correct sentences obtained after the understanding module.

To the best of our knowledge, this is the first published work that reports results obtained from field data. We have reported results for different experiments, comparing offline data, simulated task data, and field data for the five different tasks addressed. Our results show lower performance for English compared with Spanish due to the lesser amount of training data used for English and the less experience we have with this language. As regards the influence of the task in the recognition error rate, we confirm that for our experiments on ATC speech, perplexity is a good prediction variable.

The comparison of field data results with simulated data results recorded in the laboratory for one of the tasks (read data or free speech) demonstrates that the spontaneity found in field data decreases the performance of our original system mainly due to the naturalness and style of the interactions. From a 96% word accuracy for read speech it goes down to 86% for real recordings (offline) for the clearances task for Spanish. Consequently, it is not possible to know the performance of a system without testing it with field data. However we have demonstrated that a good estimation of performance can be made by processing the sentences obtained in a simulated scenario in which a professional user performs

a task and is free to use whatever sentences he wants.

Finally a discussion on application perspectives in this area has also been given. Some commercial systems are already on the market for training ATC controllers based on speech recognition, although they are only operative in constrained fields and we have no evaluation surveys from customers. With the availability of more field data we think that there is a great potential for the future development and use of speech recognition and understanding algorithms applied to ATC speech, some of which are presented in the paper.

## ACKNOWLEDGMENTS

We express our gratitude to AENA who entrusted the project to us, and 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. We would also like thank to Mark Hallett for his English revision.

## REFERENCES

[1] Revis, M.
The case for speech recognition.
*For The Record*, **17** 9 (2005), 20.

[2] Grasso, M. A.
The long-term adoption of speech recognition in medical applications.
In *Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems* (CBMS'03), 2003.

[3] Koester, H. H.
User performance with speech recognition: A literature review.
*Assistive Technology*, **13**, 2 (2001), 116–130.

[4] Raux, A., et al.
Let's go public! Taking a spoken dialog system to the real world.
Presented at Interspeech 2005 (Eurospeech), Lisbon, Portugal, 2005.

[5] Schalk, T.
VUIs in vehicles: Meeting customer expectations.
*SpeechTech Magazine*, 2006, retrieved from http://www.speechtechmag.com/Articles/Column∼The-View-from-AVIOS VUIs-in-Vehicles-Meeting-Customer-Expectations-30076.aspx.

[6] Lalande, S. and Harbluk, J. L.
Speech-based interfaces for use while driving: The impact of email speech output.
In *Proceedings of CMRSC-XIV*, Ottawa, Ontario, Canada, June 27–30, 2004.

[7] Bergl, V., et al.
CarDialer—multimodal in vehicle cellphone control application.
Presented at the 8th International Conference on Multimodal Interfaces (ICMI'06), Banff, Canada, Nov. 2–4, 2006.

[8] Ehsani, F. and Knodt, E.
Speech technology in computer-aided language learning: Strengths and limitations of a n new call paradigm.
*Language Learning and Technology* **2**, 1 (1998), 45–60.

[9] Hincks, R.
Speech recognition for language teaching and evaluating: A study of existing commercial products.
In *Proceedings of Interspeech* (ICSLP), 2002.

[10] Neri, A., Cucchiarini, C., and Strik, W.
Automatic speech recognition for second language learning: How and why it actually works.
In *Proceedings of the 15th International Congress on Phonetics Sciences* (ICPhS), Barcelona, Spain, 2003, 1157–1160.

[11] Grady, M. W. and Herscher, M. B.
Advanced speech technology applied to problems of air traffic control.
*NAECON 75*, Record 541,1996.

[12] Beek, B., Neuberg, E. P., and Hodge, D. C.
An assessment of the technology of automatic speech recognition for military applications.
*IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-25**, 4 (Aug. 1977).

[13] Weinstein, C. J.
Opportunities for advanced speech processing in military computer-based systems.
In *Proceedings of the Workshop on Speech and Natural Language*, Hidden Valley, PA, June 24–27, 1990.

[14] Eyferth, K., Niessen, C., and Spaeth, O.
A model of air traffic controllers' conflict detection and conflict resolution.
*Aerospace Science and Technology*, **7** (2003), 409–416.

[15] Taylor, G., Miller, J., and Maddox, J.
Automating simulation-based air traffic control.
*Interservice/Industry Training, Simulation, and Education Conference* (I/ITSEC), 2005.

[16] Breaux, R., Blind, M., and Lynchard, R.
Voice technology in Navy training systems.
*AGARD Lecture Series No. 129 on Speech Processing*, June 1983.

[17] Harrison, J. A., et al.
Machine supported voice dialogue used in training air traffic controllers.
*Proceedings of the IEE International Conference on Speech Input/Output: Techniques and Applications*, 258, Mar. 1986, 110–115.

[18] Hobbs, G. R.
The application of speech input/output to training simulators.
In *Proceedings of the IFS Conference on Speech Technology*, Brighton, UK, Oct. 1984, 121–134.

[19] Slemon, G. and Eames, D.
Speech technology applied to operational air traffic controller training systems.
*Proceedings of Military Speech Tech* (Media Dimensions), Oct. 1986.

[20] Hering, H.
Comparative experiments with speech recognizers for ATC simulations.
Eurocontrol Experimental Centre, EEC Note 9/98, Bretigny, France, 1998.

[21] Marque, F., et al.
PAROLE: A vocal dialogue system for air traffic control training.
*ESCA—NATO/RSG 10 Workshop on Applications of Speech Technology*, Lautrach, Germany, Sept. 16–17, 1993.

[22] Marque, F. and Neel, F.
PAROLE. Aide à la formation et l'entraînement des contr¡3d01¿leurs de trafic aérien.
*6th Aerospace Medical Panel Meeting, Symposium on Virtual Interfaces: Research and Applications*, Lisbon, Portugal, Oct. 18–22, 1993.

[23] Schäfer, D.
Context-sensitive speech recognition in the air traffic control simulation.
Universität Der Bundeswehr Munchen Fakultät Fur Luft-Und Raumfahrttechnik, Ph.D. Thesis, 2001, and *Eurocontrol Experimental Centre, EEC Note 02/2001.*

[24] Fernández, F., et al.
Automatic understanding of ATC speech.
*IEEE A&E Systems Magazine*, (Oct. 2006), 12–17.

[25] Rankin, J. and Mattson, P.
Controller interface for controller-pilot data link communications.
*Proceedings of the 16th Digital Avionics Systems Conference*, Oct. 1997.

[26] Ferreiros, J., et al.
New word-level and sentence-level confidence scoring using graph theory calculus and its evaluation on speech understanding.
Presented at Interspeech 2005.

[27] Unknown
Speech recognition trains next-generation of air traffic controllers; Nuance and Newfound communications deliver solution for Adacel MaxSim air traffic control simulator.
*Business Wire*, (July 30, 2002), http://findarticles.com/p/articles/mi_m0EIN/is_2002_July_30/ai_89845069.

[28] Abbott, M. G.
The use of speech technology to enhance the handling of electronic flight progress strips in an air traffic control environment.
In *Proceedings of Voice Systems Worldwide*, London (Media Dimensions), May 1990, 126–134.

[29] Ragnarsdottir, M. D., Waage, H., and Hvannberg, E.
Language technology in air traffic control.
In *Proceedings of the Digital Avionics Systems Conference*, 2003, 2.E-2.1,2.E-2.13.

[30] Cardosi, K., Falzarano, P., and Ham, S.
Pilot-controller communication errors: An analysis of aviation safety reporting system (ASRS) report.
U.S. Department of Transportation, Federal Aviation Administration, DOT/FAA/AR-98/17, 1998.

[31] Manning, C., Fox, C., and Pfleiderer, E.
Relationships between measures of air traffic controller voice communications, taskload, and traffic complexity.
Presented at the 5th USA/Europe Air Traffic Management R&D Seminar, 2003.

[32] Manning, C. A., et al.
Using air traffic control taskload measures and communication events to predict subjective workloads.
FAA Office of Aerospace Medicine, Washington, D.C., DOT/FAA/AM-02/4, 2002.

[33] Johnson, S.
Describe what is meant by the term "keyword spotting" and describe the techniques used to implement such a recognition system.
MPhil Computer Speech and Language Processing Speech Recognition Essay, Cambridge University, 1997.

[34] Duke, E. L., Vanderpool, C. C., and Duke, W. C.
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*, vol. 2, 1675–1690.

[35] Ferreiros, J. and Pardo, J. M.
Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations.
*Speech Communication*, **29** (Sept. 1999), 65–76, ISSN: 0167-6393.

[36] Huang, X. H., Acero, A., and Hon, H.
*Spoken Language Processing.*
Upper Saddle River, NJ: Prentice-Hall, 2001.

[37] Fernández, F., et al.
Language identification techniques based on full recognition in an air traffic control task.
In *Proceedings of Interspeech* (ICSLP), 2004, II-1565–1568.

[38] Ma, B., et al.
Multilingual speech recognition with language identification.
In *Proceedings of Interspeech* (ICSLP), 2002, 505–508.

[39] Zissman, M. A.
Comparison of four approaches to automatic language identification of telephone speech.
*IEEE Transactions on Speech and Audio Processing*, **4**, 1 (1996), 31–44.

[40] Torres-Carrasquillo, P. A., et al.
Approaches to language identification using Gaussian mixture models and shifted delta cepstral features.
In *Proceedings of Interspeech* (ICSLP), Denver, CO, 2002.

[41] Campbell, W. M., et al.
Support vector machines for speaker and language recognition.
*Computer Speech and Language*, **20**, 2–3 (Apr. 2006), 210–229.

[42] Castaldo, F., et al.
Compensation of nuisance factors for speaker and language recognition.
*IEEE Transactions on Audio, Speech and Language Processing*, **15** (2007).

[43] Ward, W. and Issar, S.
Recent improvements in the CMU spoken language understanding system.
In *Proceedings of the Workshop on Human Language Technology*, 1994, 213–216.

[44] Lechner, A., Mattson, P., and Ecker, K.
Voice recognition: Software solutions in real-time ATC workstations.
*IEEE AESS Systems Magazine*, (Nov. 2002), 11–15.

**José M. Pardo** (M'84—SM'04) earned his telecommunication engineering degree and Ph.D. both from Universidad Politécnica de Madrid in 1978 and 1981, respectively.

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. He was a Fulbright Scholar at MIT in 1983–1984, a visiting scientist at SRI International in 1986, and a visiting fellow at the International Computer Science Institute in 2005–2006.

Dr. Pardo won a National Award in 1980 for the best graduate in Telecommunication engineering and a National Award for the Best Ph.D. Thesis in 1982. He was a member of the ISCA Advisory Council from 1996 until 2006. He was chairman of EUROSPEECH 1995 and member of ELSNET Executive Board 1998–2004. He was a member of NATO RSG 10 and IST 3 from 1994 to 2002. He is a member of ASA, ISCA and EURASIP.

**Javier Ferreiros** (M'02—SM'10) earned his telecommunication engineering degree and Ph.D. 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, where he has been an associate professor since 2001. He was associate director in the Electronic Engineering Department from 2004–2008 and he is currently Director of Academic Planning of the Telecommunication Engineering School of the Universidad Politécnica de Madrid. He was a visiting scientist at the International Computer Science Institute in 1999–2000.

Dr. Ferreiros was the Technical Program Manager of EUROSPEECH 1995. He is member of ISCA. He has authored or coauthored more than 100 papers and holds one patent.

**F. Fernández** received his M.S.E.E. (2002) and Ph.D. (2008) degrees from Universidad Politécnica de Madrid (UPM), both with highest distinctions.

Since 2002 Fernando has been member of the Speech Technology Group at UPM, where he is currently an assistant professor. During 2006, Fernando stayed as a Ph.D. student at The IDIAP Research Institute affiliated with the "Ecole Polytechnique Fédérale de Lausanne" (EPFL) and the University of Geneva (Switzerland). His research interests focus on multimodal spoken dialogue systems and social signal processing.

Dr. Martínez is member of the EUCogII network.

**Valentin Sama** earned his Italian Philology Degree from the Universidad Complutense de Madrid in 1999. He is finishing the Degree of Linguistics from the same university. Since 2001 he has worked at the Speech Technology Group of the Universidad Politécnica de Madrid. From 2005 to 2008 he worked at the Natural Language Processing and Information Retrieval Group from the Universidad Nacional de Educación a Distancia and since 2008 he has worked at the Center Students with Disabilities at the same university.

Dr. Sama-Rojo has authored or coauthored around 25 papers.

**Ricardo Córdoba** (M'00) earned his Telecommunication Engineering Degree and Ph.D. degrees from Universidad Politécnica de Madrid (UPM) in 1991 and 1995, respectively.

He is a member of the Speech Technology Group since 1990, teaching in the UPM since 1993 and working as associate professor since 2003. He is Associate Director of the Electronic Engineering Department since 2008. He worked as research associate at Cambridge University, UK, Speech, Vision and Robotics Group, in 2001. His main topics of work are Dialogue systems & multimodality, language and speaker identification, and speech recognition and synthesis.

Dr. Cordoba authored or coauthored around 80 papers.

**Javier Macias-Guarasa** (M'91) earned his Telecommunication Engineering Degree and Ph.D. both from Universidad Politécnica de Madrid in 1992 and 2001, respectively.

From 1990 to 2007 he was a member of the Speech Technology Group in the Department of Electronic Engineering at the Universidad Politécnica de Madrid, Spain, where he held different teaching positions. Currently, he is associate professor in the Department of Electronics at the University of Alcala, Spain. He was a visiting scientist at the International Computer Science Institute in 2003. His main research interests are related to speech processing and the use of audiovisual sensor fusion strategies for advanced human-machine interaction in intelligent environments.

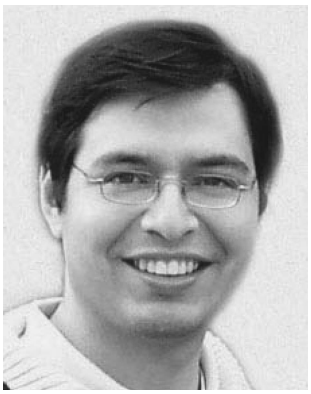
Dr. Macias-Guarasa was a member of the organizing committee of EUROSPEECH 1995. He has authored or coauthored 98 papers.

**Juan M. Montero** (M'00) earned his Telecommunication Engineering Degree and Ph.D. both from Universidad Politécnica de Madrid in 1992 and 2003, respectively. Since 1991 he has worked in speech technology and has held different teaching and research positions at the Universidad Politécnica de Madrid. He was a visiting fellow at the International Computer Science Institute in 2005 and at DFKI Saarbrucken in 2006.

Dr. Montero won a National Award for the Best PhD Thesis in 2005. He has authored or coauthored more than 60 papers and has participated in more than 40 research projects in speech and language technology.

**Rubén San-Segundo** received his M.S.E.E. and Ph.D. degrees (with highest honours) in 1997 and 2002 both from Universidad Politécnica of Madrid (UPM). From 1997 through 2001, he was supported under a Ministry of Education Grant at the Speech Technology Group (GTH), in which his research included speech recognition of spelled words, confidence measures estimation for speech recognition and understanding, and dialogue design for speech applications. During 1999 and 2000, he did two summer stays at The Center of Spoken Language Research (CSLR) at the University of Colorado, Boulder, as a visiting student. From September 2001 through February 2003, he worked at the Speech Technology Group of the Telefónica I+D. From March 2003 through March 2004, he worked at the Department of Electronic Systems and Control at the EUITT of UPM. Currently, he is associate professor in the Department of Electronic Engineering at ETSIT of UPM and he is member of the Speech Technology Group (GTH) at UPM.

Dr. San-Segundo is the Coordinator of the Spanish Network on Speech Technologies and he has been Vice-Chair of the Special Interest Group of ISCA on Iberian Languages. He has authored or coauthored more than 60 papers and has participated in more than 20 research projects in speech and language technology.



**Luis Fernando D'Haro** earned his degree as electronics engineer in 2000, from Universidad Autónoma de Occidente in Cali, Colombia, and Ph.D. (with highest honours) from the Universidad Politécnica de Madrid in 2009.

He is currently an assistant professor at UPM, Spain, since 2007. In 2005 he stayed at Computer Science VI, RWTH Aachen University, Germany, working in machine translation and language modeling, and in 2006 at AT&T labs research in Florham Park, NJ, working in multimodal dialogue interaction and interfaces. His main research interests are related to language modeling and identification, dialogue design, and data mining.

Dr. D'Haro is member of ISCA. He has authored or coauthored more than 30 papers and holds one patent with AT&T.

**Germán González** earned his Telecommunication Engineering Degree from Universidad Politécnica de Madrid in 1990.

After a few years (1988–1991) working for the military electronics industry in radio communications technology, in 1991 he joined the engineering and consultancy company ISDEFE, where he has worked in Military Air Command and Control Systems and Civil Air Traffic Control and Management Systems in Spain and Central America. On behalf of the Spanish Air Navigation Services provider AENA, he has been member of several EUROCAE working groups for the development of Air-Ground Datalink standards and he is currently member of the EUROCONTROL Link2000+ Programme Steering Group, for the operational implementation of Air Ground Datalink services in the European Continental Air Space.