Automatic Understanding of ATC Speech

F. Fernández, J. Ferreiros, J.M. Pardo, V. Sama, R. de Córdoba,
J. Macías-Guarasa, J.M. Montero, R. San Segundo, L.F. d'Haro
Universidad Politècnica de Madrid
M. Santamaría
Aeropuertos Espanoles y Navegacion Aerea
&
G. González
ISDEFE

ABSTRACT

In this paper we make a critical revision of the state-of-the-art in automatic speech processing as applied to Air Traffic Control. We present the development of a new ATC speech understanding system comparing its performance and advantages to previously published experiences. The system has innovative solutions such as detecting the Air/Ground language spoken by Air Traffic Controllers in an international airport with two official languages and the ability to adapt to new situations by automatically learning stochastic grammars from data, eliminating the need to write expensive and eternally incomplete grammars. A relevant new feature is the use of a speech understanding module able to extract semantically relevant information from the transcription of the sentences delivered by the speech recognizers. Two main assessment objectives are pursued and discussed throughout the paper: the effects of human spontaneity and the lack of linguistic coverage in understanding performance. The potential of this technology, ways of improvement, and proposals for the future are also presented.

INTRODUCTION

There have been several attempts at applying current speech processing capabilities to the problem of automatically understanding ATC speech. There are many applications such as being able automatically to store, deliver, and process the information transferred between controllers and pilots minimizing the need for manual controller actions on the ATC system, subsequently increasing the safety of operations and airport capacity by allowing the controllers to concentrate on the traffic situation in their areas of responsibility.

But the truth is that we are unable to cite really successful examples. In this paper we show that, among other causes, natural human spontaneity and lack of linguistic coverage are two factors which have a huge impact on the performance of speech understanding systems when applied to ATC. They are related because the lack of proper response to human spontaneity can be considered a problem of linguistic coverage in a sense, and, in fact, both could have the same solution if enough training data were available. We have observed and analyzed deviations from the official phraseology attained both in the vocabulary and the syntax from the real life operation of controllers. We have also explored the effect of low linguistic coverage by experimenting with our system under two different conditions whereas it had only been well trained for one. We also discuss how stochastic approaches serve to smooth out some of these problems that would be severe if strict grammars were used.

Speech technology has evolved rapidly in the last decade. It is now possible to make speech recognition systems work for a diversity of speaker identities and environmental situations (noise conditions, limited bandwidth, large vocabularies) with sensible performance, although some error rate cannot be avoided. Nowadays it is also possible not only to transcribe speech, but also understand what is said through further processing of the textual sentence obtained by the recognizer in a way that an action or a decision can be taken.

As a reference, in November 2004 in DARPA official speech recognition tests, a 16% word error rate was achieved on a very difficult English conversational telephone speech task with a vocabulary of 61K words, consuming 18 times real-time CPU on a Pentium (R) 3.4 GHz. processor.

Author's Current Address:
F. Fernandez, J. Ferreiros, J.M.Pardo, V. Sama, R. de Cordoba, J. Macías-Guarasa, J.M. Montero, R. San Segundo and L.F. d'Haro, Affiliation: Grupo de Tecnologia del Habla, Universidad Politècnica de Madrid; DIR: E.T.S.I. Telecomunicacion, Universidad Politècnica de Madrid, Ciudad Universitaria s/n, 28040-Madrid (Spain); M. Santamaria, Affiliation: AENA, Aeropuertos Espanoles y Navegacion Aerea, DIR: Div. Automatizacion, C/ Josefa Valcarcel, 30, Despacho 2.5 28027-Madrid (Spain).; and G. Gonzalez, Affiliation: ISDEFE (on behalf of AENA), DIR: Div. Automatizacion, C/ Josefa Valcarcel, 30, Despacho 2.05 28027-Madrid (Spain).

Manuscript received May 14, 2006.

0885/8985/06/ $17.00 © 2006 IEEE

However, it is very difficult to predict the performance of a speech recognition system in a real situation based on results from standard tests. The management of the unavoidable error conditions is essential to the quality perception of the users. For instance, when possible, using confidence measures on the speech recognizer, the system is able to predict an error and consequently act, maybe by just asking for a repetition or following more elaborate correction techniques. These are strategies that humans follow when they are not able to fully understand what is said.

In 1993, a pilot project was developed at LIMSI (CNRS-France) to train air traffic controllers in their tasks by using speech recognition and synthesis, creating a so-called automatic "pseudo-pilot." At that time, the recognition accuracy and the speaker dependency were considered the main obstacles to putting the system into effect.

In Hering 1998, a study compared three commercially available speech recognizers using recordings of the communication between the controller and the pilot during simulations at the Eurocontrol ATC simulation facility. The objective, similar to the previous one, was to facilitate the task of a human pseudo-pilot or automate it, in an ATC simulation task. Since the spoken sentences often included words not found in the recognition vocabulary, utterances with errors combined with aborted or interrupted sequences, and even a few segments in a different language (French), the speech samples form what Hering describes as "worst-case conditions" for the recognizers. The study aimed at the installation of a central speech recognition system in a simulation network, consequently, microphone-independent automatic speech recognition systems that use the limited frequency range of standard telecommunications facilities were chosen. The recognition rates were accordingly very poor, averaging between 26 and 39 percent word accuracy.

In 1997, a pilot project was developed to integrate speech recognition into a C-CAST system (Controller Communication and Situation Awareness Terminal) which was able to transmit, display, and receive clearances in an aircraft through a data link channel . The aim of the system was the translation of the speech from the air traffic controller into text that would be sent to the pilot through the data link channel. This initial system had several limitations, particularly the long enrollment time needed to create speech profiles for every new user as well as the operating system compatibility limitations. In 1999, the same group, after careful consideration, chose a new speech recognition engine to replace the original. This second speech recognition system had several significant improvements. The recognition engine supported strict grammar files and pronunciation variations so that the need to create speech profiles for every user was minimized. By using a grammar file, multiple messages could be understood while considering only a dictionary of the words that make up the message and without the need for every user to speak all the sentences. It also allowed the programming of pronunciation variations for the words in the dictionary, so two individually different pronunciations could be matched to the same word. However, the system showed up several significant limitations. While using a grammar file enabled a lot of flexibility and accuracy to the system compared to the initial one, the creation, refinement, and maintenance of the grammar file was one of the more difficult aspects of implementing this kind of system. The grammar file had to contain all possible phrases and commands that might be uttered by the user. The terminology and layout of the messages had to be rigidly defined and strictly adhered to by the users, an aim not possible to achieve in real life.

Finally in Schäfer 2001, several experiments were designed to study the effect that the use of a context-sensitive syntax has on the recognition performance, compared to that of a global syntax. The experimental environment comprised an en-route air traffic control simulator with a Commercial-Off-the-Shelf (COTS) speech recognition and a speech synthesis interface. The work demonstrated that the performance of automatic speech recognition systems in the air traffic control simulation can be improved considerably when a context-sensitive syntax is used. Compared to traditional, non context-sensitive speech recognizers, the recognition error rate could be reduced by about 50 percent. In contrast, there was a lot of work needed to define the context-sensitive syntax. No data was given concerning the number of words and perplexity of the language used, so the results are not easily comparable.

In 2001 we started the INVOCA project (Vocal Interfaces for Air Traffic Control) in cooperation with AENA (Spanish Airports and Air Navigation) with the Universidad Politécnica de Madrid. This was an exploratory project aimed at researching the strengths and weaknesses of speech understanding applied to ATC. The project dealt with two possible applications: a speech interface for command and control of an en-route and TMA ATC workstation as an additional input mechanism (they were also using a touch-screen interface) and an automatic understanding system to process live speech of a tower controller in a real controller-pilot communication radio channel to assess the capabilities of the system to transcribe and understand it, eventually extracting the key information from the sentences in a useful output format.

We will focus on the second application because it is the more challenging and technologically demanding. In the experience addressed herein, a new speech understanding module processes the output of the recognizers so as to obtain a semantic frame as the overall output of the whole system. These frames are made up of a variable number of attribute-value pairs formatted in an easily usable way by the ATC information servers (responsible for the saving and transferring of information regarding the actual course of the flight plan through the many systems keeping track).

The limitation of previously published experiences on the need to generate and use inflexible grammar files was overcome by the use of stochastic language models automatically learnt from application data. We processed thousands of real recorded utterances of communications between controllers and pilots. By transcribing them into text we could create a stochastic language model adapted to the task. The advantage of this grammar is that it covers not only the standard defined protocol sentences but any slight or individual syntactic variation that the controllers may use in their day-to-day communication. The system is capable of managing some new syntactic variations without error even if these variations were never pronounced in the recorded database. It is much more robust than other systems, because an official grammar mismatch does not necessarily imply an understanding error.

We prepared five systems specifically trained for each of the five different tower control positions that were operative at Madrid Barajas airport at the time of the study: arrivals, departure clearance, and take-offs plus two surface control positions: north and south. As Madrid Barajas is an international airport, both Spanish and English are official and common languages for the application and the systems had to be able to process sentences in both languages. Because of the project dimension restrictions, we did not dedicate the same effort to the five positions nor to the two languages. Most of the effort went into obtaining a sensible system for the departure clearance task in both languages, although we recorded and processed more data for Spanish than for English (for example, 7.1 hours of speech (4026 sentences) were used to train acoustic models of the recognizer for Spanish and 4.7 hours of speech were used for English (2200 sentences)). For similar reasons, we have the most cross-comparable evaluation data for the departure clearance position and this is the task on which we will center our discourse.

ARCHITECTURE OF THE SYSTEM

The system is made up of the modules shown in Figure 1.

A front-end module analyzes the activity in the input signal to estimate the beginning and ending points of a sentence and to extract features relevant to speech recognition (LPC-Cepstral coefficients, with CMN (Cepstral mean normalization) and CVN (Cepstral variance normalization)) for this sentence.

Next, two speech recognizers work in parallel, one for Spanish and the other for English. We have used our own in-house, continuous speech recognizer, with HMM (Hidden Markov Models) for context-dependent generalized triphones with 1500 states and 8 mixtures per state (Spanish) and 900 states, 8 mixtures per state (English). The search is driven by a stochastic bigram language model that assigns a score to each sequence of two words. These scores are learnt by processing text transcribed from actual controller sentences in the development phase. 4535 sentences were used to train the Spanish bigrams and 2703 for the English. The estimated test set perplexity of the task (the entropy of the language model measured on a scale that closely resembles the average number of choices the recognizer has to choose from, based only on this model) is 15.2 for Spanish and 23.2 for English. This lower perplexity evaluates an interesting restriction or helps in the recognition process if we compare it with the some 1000 words in the Spanish vocabulary. Without any language model, the recognizer would have to pick one out of these 1000 words every time a connection between words could occur. With the language model it has to select only one in 15.2, on average. This is an indication of the power of the language model we have chosen. Although it is a stochastic model that will not reject any combination of words, (something that will be essential for robustness, as we will discuss further) the fact that the probabilities are higher for well-formed sentences results in this large drop in uncertainty. Several pruning techniques allow our system to search through only about 17% of the hypothetical full search space and respond in real-time (0.63 times real time for the largest Spanish clearances task on an AMD Athlon™ XP 1800+ with 1,5G RAM). The Spanish vocabulary contains 1104 words plus 14 word-like units that we call extra-lexical units because they are models for non-lexical acoustic events (like silences, lips noise, speaker noises, hesitations like "hum," "eh," "mm," etc.) that do not follow grammar rules in their occurrence probability. In these 1104 words there are also some variants for 86 words constituting a convenient multiple pronunciation technique. Each word, even with the same grammatical identity, may have two or more different entries in the dictionary compiling different alternative pronunciations. Also, for 52 of the entries in the vocabulary, we do not have any language model because they did not appear in the training material used in the design of the system. They are given an intermediate score (the average between the largest and the shortest values in the language model) when they intervene in a sentence. In the English case the vocabulary contains 793 entries plus 14 extra-lexical units. 122 entries correspond to multiple pronunciations and 36 words are new with respect to the training and adopt the aforementioned intermediate language model score.

After going through the recognizers, the next module compares the overall scores obtained by both recognizers and chooses the best output and, thus, it determines the most probable language for the sentence. This language identification technique is more robust (yet more time consuming) than other standard approaches seen in the literature. We need this more elaborate approach because the characteristics of this task make it particularly difficult as the controllers are non-native in English. 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 day) in Spanish while the rest of the phrase is pronounced in English. The language identification error rate obtained in our experiments is 5%. It was considered reasonable for such a difficult task although it poses a significant upper limit for the overall system performance.

The output text in the language chosen is passed on to the understanding module that will extract variable length frames containing a set of attribute-value pairs as the final output. This

Table 1. Experimental results for different evaluation settings

		Recognition		Understanding	
Experiment	**Language**	**Word Accuracy**	**Perfect Sentences**	**Concept Accuracy**	**Perfect Sentences**
Simulation Guided Sentences	**Spanish**	**96.73%**	**54.29%**	**92.36%**	**68.57%**
	English	**91.42%**	**19.05%**	**83.94%**	**50.00%**
Simulation Free Sentences	**Spanish**	**89.05%**	**18.57%**	**81.77%**	**44.29%**
	English	**79.45%**	**11.90%**	**66.32%**	**21.43%**
Tower South config (Worst Case)	**Both Mixed**	**77.99%**	**17.14%**	**51.59%**	**29.17%**
Tower North Config (Best Case)	**Both Mixed**	**88.96%**	**35.61%**	**76.87%**	**52.38%**

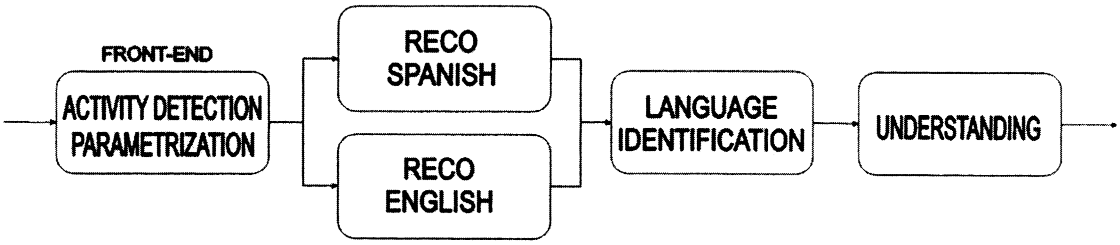

Fig. 1. Architecture of the speech understanding system

module uses context-dependent rules on the set of semantic-pragmatic labels given to each word in the recognized sentence; its operative: 99% language independent.

EVALUATION OF THE SYSTEM

To assess the effect of expression spontaneity and linguistic coverage, we will present several experiments. We will begin by classifying them into two main blocks (see Table 1).

The first block includes results from a simulated ATC departure clearance task. Each controller was given a scenario where they had to deliver 10 Spanish instructions and 6 English to fictitious pilots. In each case, the clearance was given twice: first (labeled in Table 1 as "Simulation Free Sentences"), freely generating a sentence by themselves giving the data within the framework of the scenario and second, ("Simulation, Guided Sentences"), reading a sentence that we display on screen exactly in this second phase (extracted from the set of live recordings used during system development and with the same semantic content). We got the help of 7 student controllers from SENASA ("Sociedad para las ENseñanzas Aeronáuticas civiles, S.A.," a Spanish controller training institute) for this experiment. This block was designed in order to isolate the effect of natural human spontaneity on understanding ATC commands.

The second block contains experiments using the complete definitive understanding system directly connected to a live departure clearance radio channel at Madrid Barajas international airport. By chance, on one of the two experimentation days we found the airport in a configuration for south winds ("South Configuration") instead of the more usual north configuration to which all the training material we captured at the beginning of the project belonged. This circumstance caused a lack of linguistic coverage that allows us to discuss its isolated effect.

Table 1 contains the following figures:

- *Word or concept accuracy,* both calculated as complementary to the total error rate (the first, for the words output by recognizer and the second, for the concepts from the ontology of the application output by the final understanding module). The total error rate includes deletion, substitution, and insertion error rates added together. In the case of the understanding stage, we decided to count a substitution when caused both because of a substitution of the attribute or of the value assigned to the correct attribute. This constitutes a very strict and conservative performance measurement.

- *Perfect sentences,* which means, for the speech recognizer, that the sentence has been perfectly transcribed word for word and, for the understanding stage, that all the attributes and their given values match the expected ones.

The first thing that can be appreciated just by comparing, for any experiment, the columns of perfect sentences for the recognition and the understanding stages, is a huge increase in the number of perfectly processed sentences by the understanding module compared to those perfectly transcribed by the recognizer. This is in fact an expected feature of the understanding process since it does not need the perfect transcription of all words to produce a correct interpretation of the sentence. If the errors fall on semantically irrelevant words or on parts of the sentence with semantic redundancy in another part, the understanding module is able to do its job equally well.

A second analysis will be derived from the observation of the better performance obtained from Spanish systems compared to English systems (in both "Simulation" experiments), both at the word recognition and final understanding output levels. There are multiple causes for this effect: we had more Spanish data in the project recordings; we have more experience in building Spanish systems and more knowledge about the language that influences our capacity for deciding on optimal phone inventories, multiple pronunciations, etc.; and finally, English examples are uttered by non-native speakers with a very high variability in pronunciation.

After these two general observations, we will go deeper into the assessment that constituted the objective of this work. First, we can check the drop in performance in the "Simulation" experiments caused by the effect of natural human spontaneity looking at the results labeled "Free sentences," freely elaborated for the given scenario by the controllers before knowing the sentence they were also required to read later, and comparing these to those labeled "Guided sentences," the read sentences. In all the cases (recognition and understanding for both Spanish and English) we observe this significant drop in performance, even though when studying the experiment we found very few OOVs (Out of Vocabulary Words, i.e., words not previously known by the system), 6 for the Spanish sentences and 1 for English. This highlights the conclusion that the main differences between both experiments have to be related to syntactic variations, even though we have been careful to use stochastic grammars instead of more restrictive ones. It is also true that each OOV caused two or more accumulated errors in the transcription generated by the speech recognizer in especially harmful places for language understanding purposes.

Finally, we show the results in the real experiments with the full system running on a live departure clearance radio channel at Barajas tower (second block of the table). We would like to point out that in these experiments, the language identification module (that decides whether the sentence was uttered in English or Spanish) is a decisive factor that introduces its intrinsic 5% error as an upper limit to the performance of the full system. In cases where the language is badly recognized, all efforts to understand the content are certainly wasted.

As mentioned, we have two cases. The North configuration for which our system was originally trained and for which we get the better performance. It is interesting to note that more than 52% of the sentences are understood without the slightest error in the interpretation of the contents. The other case, South configuration is the worse case and produces problems of lack of coverage that impact directly on the observed drop in performance in all figures, although 29% of the sentences are still fully understood in this odd condition for our system. Our thought is again that this remaining robustness is provided by the use of stochastic grammar models that do not reject sentences with a slight coverage problem so they can be properly processed by the understanding module. In this experiment we found 45 OOVs that contributed to the errors had a significant effect of the lack of language model coverage. Many of the OOVs were Spanish words, something that could be explained considering that the controllers are more likely a larger diversity of words in their mother tongue, leading again to a kind of spontaneity factor.

DISCUSSION

COTS systems as used in previous experiences need a specific grammar to be developed with a great deal of effort and is never complete. Out-of-grammar sentences result in big errors. Furthermore, some characteristics for improvement such as adapting acoustic models to specific speakers are not usually available. With a design customized to the task, as is done in this work, results can be much better and more robust if automatic learning techniques are used. For these stochastic schemes, the quantity of data available for training determines the resulting performance of the system that we have experimented on obtaining better results for Spanish than for

English. Experience and knowledge about the language are also relevant factors in the design of the recognition systems.

There are several ideas that could be implemented to improve the performance of our ATC speech understanding system. We could get a significant improvement just by applying pragmatic constraints. As a result of this work we have gathered some useful knowledge from the ATC domain. 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. A second improvement has to do with available training data which we have used to develop the system. We have used a reduced amount of data resulting from our time and effort limitations, so out of domain phrases have often appeared. This error source could be significantly minimized just by collecting more data. Third, training data and test data differ depending on the particular speaker. A speaker dependent acoustic and linguistic modeling could significantly improve the obtained results. Previous experiences as in Cordoba 2005 [1] show that an error reduction up to 80% is possible by carrying out speaker adaptation. Finally, if a limited workload can be allowed for the user, a confidence mechanism could be implemented in order to predict a sentence error in advance and to ask for a repetition (as actually happens in human-human communications). In this way, the user does not have to correct the errors by hand thus producing a positive feeling about the intelligence/performance of the system.

CONCLUSIONS

From all results obtained, we have been able to analyze the current power of speech recognition technology applied to air traffic control. Results may not seem good enough for a definitive integration of our system into the tasks of the controller (integration into a real operational ATC system). Nonetheless, it is very important to emphasize that these results must be considered as a first approximation since recognition rates could be significantly improved by just following some of the ideas and possibilities we have proposed but which have not yet been implemented into our current system.

New, highly-interesting potential domains exist for speech recognition systems for which current performance levels would certainly be acceptable. One example is in the field of the training of future controllers. The developed system would be perfectly suitable for this area since these are not critical systems. Despite ATC training systems having to be identical to the real operational ones, certain differences are derived from the training procedures can be assumed, as long as they do not imply any change in the controller tasks. This is the case of implementing a speech interface, which would help the automation of the training process and the trainee performance monitoring. The acceptable recognition error rate for this interface could be lower than for a hypothetical interface integrated into an operational ATC system. Even a certain level of error may be useful in order to better simulate an understanding problem with a pilot or with the communication channel. The scope of such systems would range from an automatic pseudo-pilot (that automatically reacts to instructions given by the trainee and execute simulated aircraft maneuvers) to a phraseology trainer in which the system would rate the adhesion of the ATC students to the official syntax and recommended speech procedures.

ACKNOWLEDGEMENT

We express our gratitude to the personnel in AENA and SENASA who helped us in many stages of the development of this project. We also thank J. Colas for his work in the early stages of the project.

REFERENCES

[1] R. de Cordoba, J. Macias, V. Sama, R. Barra and J.M. Pardo, (2005),
New advances in cross-task and speaker adaptation for air traffic control tasks,
Procesamiento del Lenguaje Natural, Vol. 35, pp. 21-27, September 2005.

[2] Hering, H., (1998),
Comparative Experiments with Speech Recognizers for ATC Simulations. EEC Note No. 9/98,
Eurocontrol Experimental Centre, Eurocontrol, Bretigny, France.

[3] A. Lechner, P. Mattson and K. Ecker,
Voice Recognition: Software Solutions in Real-time ATC Workstations,
IEEE AESS Systems Magazine, November 2002, pp 11-15.

[4] F.Marque, S. Bennacef, F. Neel and S.Trinh,
Parole : a vocal dialogue system for air traffic control training,
ESCA Workshop *Applications of Speech Technology,*
Lautrach, Germany, September 16-17, 1993.

[5] F. Marque and F. Neel,
PAROLE. Aide a la formation et l'entrainement des controleurs de traffic aerien,
6th Aerospace Medical Panel Meeting. Symposium Virtual Interfaces: Research and Applications,
Lisbon, Oct. 18-22, 1993.

[6] J. Rankin and P. Mattson,
Controller Interface for Controller-Pilot Data Link Communications,
Proceedings of the 16th Digital Avionics Systems Conference, October 1997.

[7] Dirk Schafer,
Context-Sensitive Speech Recognition in the Air Traffic Control Simulation,
in Universitat Der Bundeswehr Miinchen Fakultat Fur Luft- Und Raumfahrttechnik, Phd. Thesis, 2001, and
Eurocontrol Experimental Centre EEC Note No. 02/2001.