4th USA / Europe Air Traffic Management R&D Seminar Santa Fe, 3-7 December 2001

Context-Sensitive Speech Recognition in the Air Traffic Control Simulation

Dirk Schäfer

EUROCONTROL Experimental Centre
Bretigny, France
dirk.schaefer@eurocontrol.int

Abstract

Attempts to introduce automatic speech recognition systems in air traffic control (ATC) and especially in ATC simulation have been numerous and almost consistently concluded that "today's speech recognition technology just isn't good enough for this domain" [4] yet considered on the brink of being used widely. Rather than heralding a break through, serious research nowadays focuses on step-wise improvements. New approaches that target at improving the recognition performance of automatic speech recognition systems (ASR) are therefore more than welcome. This paper describes the use of contextual knowledge to enhance the performance of automatic speech recognition in the domain of air traffic control, especially ATC simulation.

Unlike their human counterpart, automatic speech recognizers do usually not posses contextual knowledge, i.e. dynamic knowledge about the topic of the conversation. However, their access to contextual knowledge is one reason for the superior performance of human 'speech recognizers'. Consequently, a promising strategy to increase the recognition performance of ASR systems would be to make use of context knowledge during the decoding process and this in fact is the approach adopted and discussed in this paper.

A cognitive model of the air traffic controller has been implemented which estimates the probability of individual instructions in the actual situation and translates these into sequences of words the air traffic controller is likely to speak in the actual situation. By use of this estimation the speech recognizer learns to discriminate between likely and less likely utterances and thus gains a very basic form of context awareness.

The Cognitive Controller Model (CCM) has been implemented and connected to a speech recognizer and an ATC simulator. Experiments have been carried out during which air traffic controllers had to control traffic in a German en-route sector. The experiments are very encouraging: a mean sentence recognition rate of 93 percent has been demonstrated under conditions that controllers rated as comparatively realistic.

Introduction

Air traffic control simulation is playing an increasingly important role both in the training of ATC controllers and in research and development since simulation permits to reproduce any desired air traffic situation without causing safety hazards to human operators and machines. As ATC simulators typically do not involve real aircraft but display the movements of artificial aircraft symbols on a radar screen, the communication between controllers and pilots is mostly simulated using the so-called pseudo pilot concept. Auxiliary personnel, denoted to as 'pseudo pilots' or 'ghost pilots', are responsible for the maneuvers of the simulated aircraft, They enter the controllers' clearances into the simulation computer and confirm the instructions via a simulated radio line. However, as the required personnel makes simulations very expensive and fairly inflexible, the introduction of automatic speech recognition (ASR) systems is considered as a promising alternative to the pseudo pilot concept.

A speech recognition system for this application must understand and execute the controller's spoken commands and generate an artificial pilot response by means of speech synthesis. In order not to affect the realism of the simulation, it must be possible to communicate with the speech recognizer in a similar way as with pilots or pseudo pilots. The ASR must understand continuously spoken ATC instructions in a speaker-independent fashion with little response time. Moreover, the recognition error rate, i.e. the percentage of incorrectly decoded utterances, must not greatly exceed that of human listeners. A sentence recognition rate of 95 percent is commonly considered as the threshold for successful use in domains such as ATC [6, 10]. Unfortunately, state-of-the-art speech recognition systems do not fulfill these requirements.

Typically a dictionary and syntax specify the sentences a speech recognizer is designed to identify. While the dictionary contains the acoustic properties of each word, the syntax specifies sequences of words that are sensible for the purpose of the application. Accordingly, the definition of the syntax is crucial for the performance of the speech recognizer: relevant sentences that are not encompassed in the syntax can not be recognized. On the other hand the probability of incorrect decodes certainly increases with the number of sentences considered during the recognition process, so that the syntax should only contain the necessary sentences.

Usually, the syntax and the dictionary are created once when the speech recognizer is integrated and adapted to the application domain. The sentences that are taken into consideration during each decoding process, therefore, remain unchanged and are independent of the context or the dialogue state. One approach to overcome this limitation is the definition of *sub-contexts*, i.e. situations in which only a limited number of responses need to be considered. While sub-context building is widely used in applications such as telephone dialogue systems or menu-based command and control systems, such a discrimination can hardly be made for dialogues in ATC.

The superior performance of human listeners is not only due to the more complex sensory apparatus that humans possess or the use of non-verbal information, such as gesture. Humans also posses a broad source of contextual knowledge that permits them to anticipate and interpret which words and sentences their opposite may possibly speak and thus to comprehend even ambiguous and partially omitted utterances. The structure of the language, the subject of the conversation, the dispositions of the conversant, the present situation and the dialogue history are pools of information that permit humans to distinguish between verbal statements that are probable or less probable in the actual context.

When compared to humans, automatic speech recognition systems use little information about the subject of the dialogue. Young et al. propose a distinction of three different categories of knowledge about a conversation [19]:

- syntactic knowledge: knowledge about the structure of grammatically correct sentences in a language,
- semantic knowledge: knowledge about sentences that are meaningful in a certain domain,
- pragmatic knowledge: knowledge about the context and the actual system state.

Speech Recognition in the Air Traffic Control Simulation

A number of studies have aimed at introducing automatic speech recognition in the domain of air traffic control. Predominantly, these studies targeted at air traffic control simulation [18, 8, 6, 9, 5, 3, 19] rather than at operational applications [4]. Traditionally these approaches used a static syntax, i.e. a syntax that does not change during the execution of the exercise. Figure 1 depicts the application of speech recognition in the air traffic control simulation, replacing the 'pseudo pilot'.

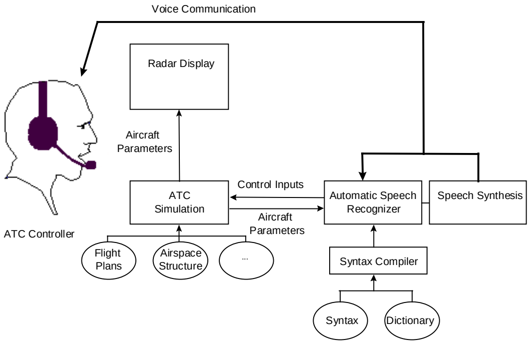

Figure 1: Speech Recognition in the ATC Simulation (static syntax).

Since the phraseology of air traffic control instructions is determined by international regulations[1] and implemented through national

---

[1] through the *International Civil Aviation Organization (ICAO)*

authorities [e.g. 1], the application of syntactic and semantic knowledge is fairly straightforward and most existing ATC speech recognizers use these sources of information.

For example, descend clearances are phrased *<callsign> DESCEND TO FLIGHT LEVEL <FL>*. Applying this syntactic knowledge already results in a significant reduction of the number of valid sentences when compared to random sequences of all words in the vocabulary. Using (semantic) knowledge about valid parameters of a descent clearance (e.g. Flight Levels in the lower airspace: between 100 and 240 in steps of 10) valid clearances are defined *<callsign> DESCEND TO FLIGHT LEVEL <100/110/…/240>*.

Using pragmatic knowledge, e.g. the aircraft's present flight level, say 140, helps to reduce valid instructions. Say its present flight level was FL140: *<callsign> DESCEND TO FLIGHT LEVEL <100/110/…/130>*. Whereas syntactic and semantic knowledge is static, pragmatic knowledge is dynamic by definition and its use for the construction of ASR search spaces therefore comparatively difficult.

One approach to enhance the performance of ASR systems would be to make use of situational information, i.e. pragmatic knowledge for a dynamic adaptation of the ASR syntax. Thus, the number of sentences in the syntax could be reduced and consequently, the recognition performance increased. This approach of context-sensitive speech recognition would require a module that, after carefully analyzing the actual situation, provides and continuously updates a prediction of sentences the air traffic controller might possibly say. This prediction could then be used to generate a dynamic, context-sensitive syntax to be used by the speech recognizer.

Firstly, the search space can be limited to sentences that are physically sensible, e.g. aircraft can only be advised to reduce their speed to a parameter less than their actual speed. Of course, this adaptation means considering the actual state of each individual aircraft and thus requires a module that calculates possible clearance parameters and transfers them to the speech recognizer.

Secondly, the controller's behavior can be described as the application of specific strategies and procedures in order to attain specific goals. Therefore, only a small fraction of the physically possible clearances are probable in each situation. A second controller looking over his colleague's shoulder is usually able to limit the focus to a number of potential strategies and instructions. If the syntax could be tailored to the actual situation, an improvement in the recognition accuracy could be expected.

The success of this approach, however, depends very much on the quality of the predictions that are used to construct the context-sensitive syntax. The restriction of the syntax to physically sensible sentences only requires an objective evaluation of the aircraft parameters and is expected to bring about some improvements in the recognition performance.

The predictive assessment of the probability of individual clearances, however, is much more demanding. The 'situation-specific' tailoring of the syntax is expected to result in significant improvements in recognition performance and this paper is dedicated to this approach. Initial attempts (by the same author) to correlate the occurrence of clearances to situation parameters in a statistic approach have not been very successful. A more promising approach seemed to be to study the mental processes involved in the controller's decision making and to implement these in a computer-based model. An excellent application of this concept in the domain of the interaction of a pilot with a fighter aircraft cockpit can be found in [12, 7].

Although the mental processes involved in air traffic control have been subject of research for many years, no existing model is capable of predicting what an air traffic controller might actually do in a specific situation. This is partly due to the enormous complexity of the tasks and cognitive activities involved in air traffic control and the difficulties in correlating user actions to the situation. It is also due to the fact that each situation can be dealt with in various manners and it is a question of expertise and individual preferences how each individual controller will actually react. A cognitive model of air traffic controllers had therefore to be developed, capable of predicting possible control actions dynamically.

CCM - A Cognitive Model of the Air Traffic Controller

To provide a continuous assessment of the most probable controller actions[2], the *Cognitive Controller Model* (CCM) has been developed. CCM is a cognitive model of the controller's activity which analyzes the traffic situation and assesses which instructions the controller may be expected to issue in that situation, quite similar to a second controller looking over his colleague's shoulder. The architecture of CCM is based on existing models of human problem solving from the areas of cognitive psychology [13, 14]. The decision processes and strategies implemented in CCM have been derived by observing and questioning ATC controllers during extensive simulations and were calibrated using data logs from simulation runs.

CCM consists of three major modules: The *observation module* is generating a mental representation of the actual context by observing and classifying the present air traffic situation. The *decision module* aims at predicting which instructions a controller could issue under the given circumstances. The *phraseology module* translates the instructions into sentences, i.e. sequences of words, and generates a context-sensitive syntax. As the controller's decisions are to a large degree dependent on the airspace geometry, the implementation of CCM is specific to the ATC sector. A German en-route control sector (Frankfurt WR1) in the vicinity of Frankfurt was selected for the first implementation. Figure 2 depicts the Cognitive Controller Model. Data material recorded during simulations in an ATC simulation facility was used in order to calibrate CCM and test its performance.

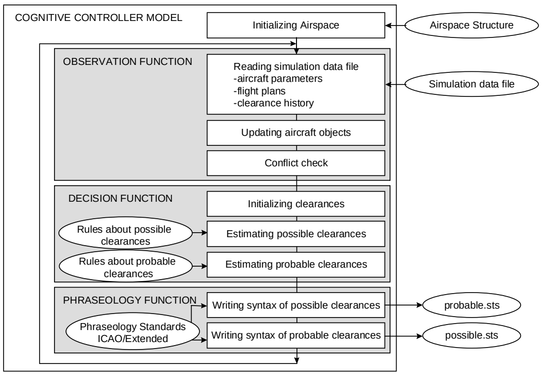

Figure 2: The Cognitive Controller Model (CCM).

The Cognitive Controller Model has been implemented as a software program A different methodology of knowledge representation was found to be adequate for the three different modules of CCM. An object-based structure was chosen for the observation module, whereas the controller strategies in the decision module were implemented as production rules. In order to generate the syntaxes, the phraseology module maps phraseology items to the more symbolic descriptions of the ATC clearances used within CCM.

CCM produces two dynamic syntaxes: the syntax containing the most probable instructions as considered by CCM *(probable)* is used as the primary means for decoding the spoken utterance. If for some reason the decode should fail, possibly because the controller decided on a strategy CCM did not anticipate, the decoding process is repeated using a syntax containing all instructions that are physically possible in the actual situation *(possible)*. Both syntaxes are generated once per second.

CCM was integrated into an ATC simulation environment equipped with a speech interface consisting of an off-the-shelf automatic speech recognizer and a speech synthesis system. The ATC simulation environment permits to control simulated aircraft by means of voice, speaking the instructions in the conventional manner and provides an artificial read-back of the instructions. The speech recognizer can be configured to use either the static syntax or the

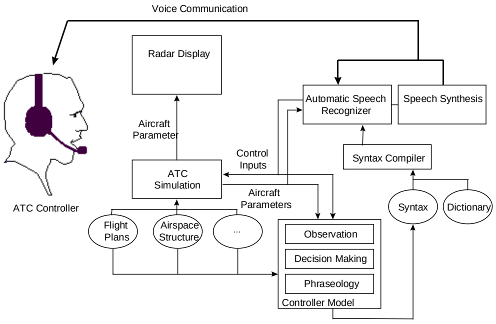

Figure 3: Context-sensitive speech recognition in the ATC simulation using CCM.

---

[2] rather than a deterministic prediction of his behavior which is not feasible for reasons given above

context-sensitive syntaxes generated by CCM. Figure 3 depicts the Cognitive Controller Model integrated into the ATC speech recognizer. A comparison to Figure 1 shows that the static syntax in conventional speech recognizers is now replaced by a dynamic syntax, in this case updated once per second.

The official phraseology standards, as defined by international regulations, were implemented in the *ICAO standard phraseology*. However, controllers often use phrases that deviate slightly from the official standards and it was presumed that when being compelled to use the standard phraseology speaking would require additional attention. This would be a perturbation of the experimental conditions and impede the transfer of simulation results to reality. Therefore, the most frequent deviations from the official standards were collected and implemented in the *extended phraseology*. CCM was implemented with a feature permitting the experimenter to choose whether the syntax shall be generated in the ICAO standard phraseology or in the extended phraseology.

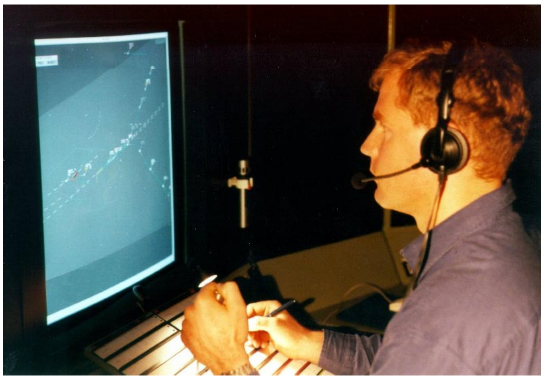

Figure 4: The speech recognizer in the ATC simulation environment.

In the ATC simulator the controller interacts with the virtual traffic in very much the same way he would work in real life (see Figure 4). Control clearances are spoken into the microphone attached to the controller's headset and a *push-to-talk* button must be pressed for the duration of the transmission. The responses of the virtual flight crews are generated by a semi-synthetic speech synthesis. This 'synthesis' is based on the concatenation of digitized sound items that are stored on the computer's hard disk drive in a sound file format. Since the pilot responses are usually structured

<callsign> <clearance> <parameter>

and the vocabulary of ATC clearances and parameters is fairly limited, it was possible to record all the required sound items. The pilot readback is then generated by concatenating the required items. E.g. if the controller has cleared the aircraft DLH1905 to descend to flight level 130:

LUFTHANSA 1905 DESCEND FLIGHT LEVEL 1-3-0

the pilot read back is generated by sequentially outputting three items:

LUFTHANSA 1905 // DESCENDING FLIGHT LEVEL // 1-3-0

The speech synthesis uses a repository of roughly 300 speech items, including aircraft callsigns and navigation aids. Compared to fully synthetic speech, digitized speech is naturally sounding and easier-to-comprehend.

Experimental Evaluation

In order to evaluate context-sensitive speech recognition experiments were conducted at the ATC academy of the German Air Navigation Services (DFS). More specifically, the aims the simulations was to quantify the effect of using the dynamic syntax generated by CCM instead of the static syntax (compare Figure 2 and Figure 4). The experiments consisted of simulations during which the test participants had to control the traffic by means of voice quite similar 'real' air traffic control. The test participants were air traffic controllers working as instructors at he ATC academy.

Experimental Design

The major research objective was to test whether context-sensitive speech recognition would result in increased recognition performance. Hypothesis A was formulated according:

HA: The use of the context-sensitive syntaxes dynamically generated by the Cognitive Controller Model results in a reduced recognition error rate, as compared to the use of the static syntax.

A second research objective was to investigate whether the use of the extended phraseology would reduce the recognition error rate:

HB: The use of the extended phraseology, compared to the use of the ICAO standard phraseology, has a negative effect on the recognition error rate.

Half of the test subjects used the ICAO standard phraseology whereas the others used the extended phraseology. The experimental design targeted at testing the counter-hypothesis HA_0 and $HB_{0:}$ HA_0 was tested via within-subject design whereas HB_0 was tested via between-subject design. The parametric Wilcoxon test was applied to test HA_0 (uni-directional) whereas the Mann-Whitney-U test was used to test HB_0 (uni-directional).

The experiment consisted in simulations in the en-route control sector Frankfurt WR1. 31 aircraft had to be controlled during 70 minutes, most of which where arrivals at Frankfurt airport. 13 licensed air traffic controllers participated in the simulations, eleven male and two female. The average experience as air traffic controller was 28 years, only one test subject was licensed for WR1.

The first simulation run helped the test subjects to familiarize themselves with the simulation environment and, more importantly, the handling of the speech recognizer. Each participant then controlled the traffic during two measured simulations, during one of which the speech recognizer used the dynamic syntax generated by CCM, whereas in the other simulation the static syntax was used. Due to hardware limitations, it was only possible to speak one instruction per transmission whereas in reality controllers often include two or more instructions in one transmission.

Different traffic samples were implemented so no test subject was confronted with the same traffic scenario twice. The experimental plan was balanced in order to exclude sequence effects (e.g. learning effects). A total of 26 simulations were analyzed for which audio recordings were transcribed and compared to the speech recognizer decode logs. In total more than 4100 sentences were analyzed. According to the contents the transmission were classified as belonging to the categories init-call/change frequency, altitude/flight level, heading, waypoint, speed, vertical speed, and other.

After the simulations, test subjects were asked to fill in questionnaires concerned with the usability of the speech recognizer and the realism of the simulations.

Results

Figure 5 compares the recognition error rates for simulation with the static syntax and those with the dynamic syntax generated by the cognitive controller model. Overall, the recognition error rate was reduced by ca. 50 percent. This was observed both if the ICAO standard phraseology or the extended phraseology was used. The reductions in the recognition error rate were stronger in some clearance categories than in others, but significant in most cases.

The recognition error was higher when the extended phraseology was used, however, this effect was statistically not significant. The experiments suggest that the recognition error rate is slightly smaller for female speakers than for male speakers, but the limited data does not permit a significant conclusion.

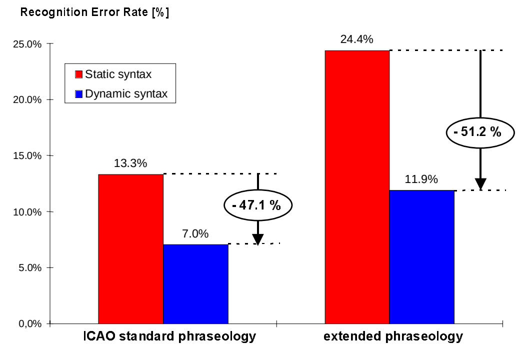

Figure 5: Mean recognition error rates.

No significant effects were found for workload ratings using the NASA TLX. Neither did the subjective workload change between simulation with static vs. dynamic syntax nor could changes be observed when the extended phraseology was used instead of the ICAO standard phraseology. Subjectively, the TLX ratings seemed to be determined primarily by the recognition error rate.

Figures 6 and 7 show the recognition error rates per clearance type, both for the ICAO standard phraseology and the extended phraseology. Remarkably, the reduction in the error rate varies greatly and is highest for clearances of type *Fix*. CCM uses the aircraft's flight plan and destination for the prediction of those fixes the aircraft may be cleared to, thus the syntax *probable* contains only a small number of fixes. Should the recognition fail, e.g. because the aircraft has been redirected to another destination, the decoding process would be repeated using the syntax possible, containing all relevant fixes.

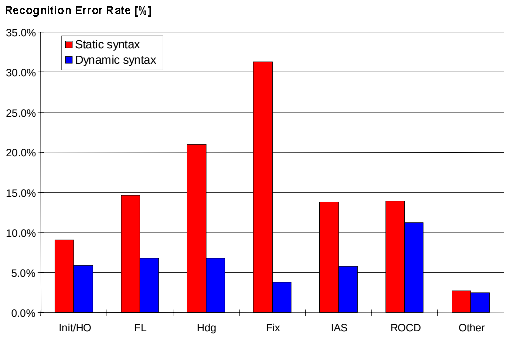

Figure 6: Error rate per clearance type - ICAO standard phraseology.

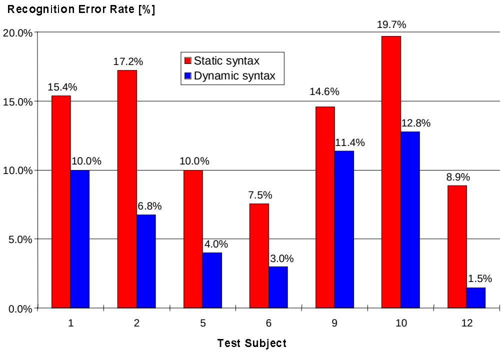

Figure 8: Error rates per test subject – ICAO standard phraseology.

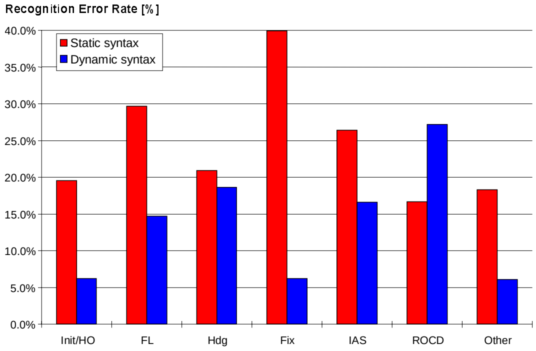

Figure 7: Error rate per clearance type – extended phraseology.

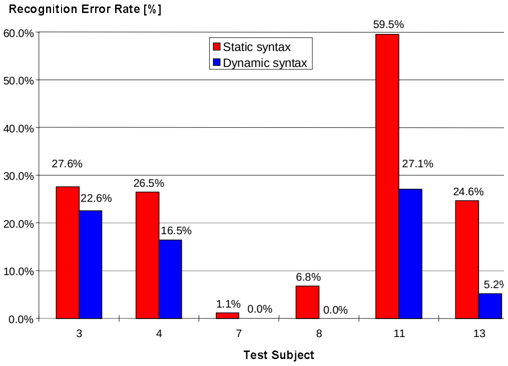

Figure 9: Error rates per test subject – extended phraseology.

Figures 8 and 9 show the recognition error rates per test subject, both for the ICAO standard phraseology and the extended phraseology. The differences in recognition performance between the individual test subjects is striking. Especially Figure 9 illustrates that the recognition error rate varies between 59.5 percent and 1.1 percent (static syntax) and between 27.1 percent and zero (dynamic syntax). Obviously, the speaking characteristics of individuals are more or less conformant to the speaker model of the speech recognizer. The test subjects, however, were all German native speakers, speaking ATC English, while the speech recognizer used a US speaker model. It may be argued that the use of a specifically adapted speaker model might reduce the differences between individuals and enhance the overall recognition rate.

Asked about the usability and the perceived effect of CCM, most test subject said the effect that CCM brought about the recognition performance was perceptible. Most participants said the simulation environment was sufficient to be used in the training of novice controller disciples. The quality of the speech synthesis and the response times were rated as agreeable.

Conclusions

Obviously, CCM has the potential of greatly enhancing the speech recognition process. Compared to traditional, non context-sensitive speech recognizers, the recognition error rate can be reduced by about 50 percent. An average recognition rate of 93 percent was demonstrated under conditions controllers rated by test participants as realistic. The response time and the handling of the speech recognizer have been judged agreeable and it seems very probable that speech recognition will play a vital role in future ATC training and simulation systems. However, controllers felt restricted by the fact that they could not speak more than one instruction per transmission. Limitations related to the hardware will probably be overcome by using new, faster computers and different speech recognizer hardware in the future.

With further experience it will be possible to continuously improve CCM and the speech recognizer, so that the performance will permit to use speech recognition in more advanced simulations. As soon as this new technology has given proof of its performance and reliability, questions about its use in operational ATC may be raised.

The knowledge base within CCM which is required for the prediction of probable clearances is dependent on the airspace sector. Since it's creation requires to study the procedures and working methods of controllers in the sector, generating this knowledge base is costly and CCM therefore quite inflexible. A broader use of the approach of context-sensitive speech recognition seems to depend on a model that is independent of the airspace sector. The prediction of physically possible clearances constitutes a first approach towards such sector-independence and could be elaborated further.

The speaker model that present speech recognizers use are typically based on native speakers, either US or British English. Non-native speakers speaking ATC English may differ from native speakers and it might be studied whether the generation of a specific speaker model would bring about performance increases.

References

[1] *Aeronautical Information Publication*, DFS Deutsche Flugsicherung GmbH, 1997.

[2] Cardosi, K. (1993): *An Analysis of En Route Controller Pilot Voice Communications*. John A. Volpe National Transportation Systems Center, MA, USA.

[3] Chung, S.; Moldovan, D. I.; DeMara, R. F. (1993): *A Parallel Computational Model for Integrated Speech and Natural Language Understanding*. IEEE Transactions on Computers, Vol. 42, No. 10.

[4] Churcher, G. (1996): *Speech Recognition for Air Traffic Control*. Technical Report, Leeds University, School of Computer Studies, Leeds, UK.

[5] David, H.; Pledger, S. (1995): *Intrusiveness of On-Line Self-Assessment in ATC Simulation Using Keyboard and Speech Recognition.* EEC Report No. 275, Eurocontrol Experimental Centre, Bretigny, France.

[6] Dunkelberger, K.; Eckert, R. (1995): *Magnavox Intent Monitoring System for ATC Applications*. Magnavox.

[7] Gerlach, M. (1996): Schnittstellengestaltung für ein Cockpitassistenzsystem unter besonderer Berücksichtigung von Spracheingabe. *VDI Fortschrittsberichte, Reihe 12, Nr. 273*, VDI Verlag, Düsseldorf, Germany.

[8] Hamel, C.; Kotick, D.; Layton, M. (1989): *Microcomputer System Integration for Air Control Training.* Special Report SR89-01, Naval Training Systems Center, Orlando, FL., USA.

[9] Hering, H. (1995): *Speech Recognition Experiments with the TRACON/Pro ATC Simulator.* EEC Note No. 2/95, Eurocontrol Experimental Centre, Eurocontrol, Bretigny, France.

[10] Karlson, J. (1989): *Automatic Speech Recognition in Air Traffic Control - A Human Factors Perspective.* Joint University Program for Air Transportation Research (NASA-CP-3095), Dec. 1990, p. 9-14.

[11] Niessen, C.; Eyferth, K.; Bierwagen, T. (1997): *Modeling cognitive processes of experienced air traffic controller*. Proceedings of the Sixth European Conference on Cognitive Science Approaches to Process Control, Baveno, Italy.

[12] Onken, R. & Gerlach, M. (1993): *Speech Input/Output as Interface Devices for Communication Between Aircraft Pilot and the Pilot Assistant System CASSY*. Joint ESCA - IEEE - NATO/RSG 10, Tutorial and Research Workshop "Application of Speech Technology".

[13] Rasmussen, J. (1983): *Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models.* IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-13, No. 3257, May/June 1983, p 257 - 266

[14] Rouse, W.B.; Rouse, S.H. (1986): *Analysis and Classification of Human Error.* IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-13, No. 4, July/August 1983, p 539 – 549

[15] Schäfer, D. (1996): A cognitive model of expert behaviour in an air traffic control task: enhanced speech recognition using situational knowledge. In: Harris, D. (Ed.): *Engineering Psychology and Cognitive Ergonomics,* Ashgate, U.K.

[16] Schäfer, D. (2001): *Context-sensitive Speech Recognition in the Air Traffic Control Simulation*, Eurocontrol Note 02/2001 (http://www.eurocontrol.fr/public/reports/eecnotes/2001/2.htm)

[17] Taylor, M. R. (1989): *Studies in Automatic Speech Recognition and its Application in Aerospace*. C.N.A.A. Polytechnic of Central London, London, UK.

[18] Weinstein, C. (1991): *Opportunities for Advanced Speech Processing in Military Computer-Based Systems.* Proceedings of the IEEE, Vol. 79, No. 11.

[19] Young, S.; Hauptmann, S.; Ward, S.; Smith, S.; Werner, S. (1989): *High level Knowledge Sources in Usable Speech Recognition Systems.* Communications of the ACM, Vol. 32, No. 2, p183-194.

Acknowledgements

This research was sponsored and carried out at the German Aerospace Center's (DLR), Institute of Flight Guidance and had been submitted as a doctoral thesis to the German Armed Forces University. The author wishes to thank Heinz Winter, Reiner Onken, Ralf Beyer, Marc Gerlach, Frank Emmermann, and Bernd Engelhard.

The experiments were conducted at the ATC academy of German Air Navigation Services (DFS).

Author Biography Dirk Schäfer

Dr. Dirk Schäfer works as a Human Factors specialist at the Eurocontrol Experimental Centre in Bretigny, France. He received an Aerospace Engineering diploma degree from Aachen Technical University in 1994 and a doctor degree from the German Armed Forces University in Munich in 2000. Between 1994 and 2000 Dirk was employed with the German Aerospace Center where his work was focussed on human factors and ATM simulation. His research focus presently lies with Human Factors aspects of Free Flights/Free Routing as well as Multi-Sector Planning.