ESCALE

Enhanced Speech Tracking of Air Traffic Control Communications

IntuiLab
Institut de Recherche en Informatique de Toulouse
Technical University of Crete
THALES Avionics
THALES Research & Technology France

Eurocontrol – CARE INO II programme

1. Introduction

The reliability of the communications between air traffic controllers and aircraft pilots is of paramount importance to air traffic safety. As such, the detection of communication problems between controllers and pilots has always been a major safety issue. Many of such problems arise from undetected misunderstandings during radio conversations. An automated tracking of the controller-pilot dialogue, in order to check the matching between clearances (controller) and acknowledgments (pilot), coupled with a verification of the effectiveness of the corresponding modification of the flight parameters, would dramatically improve flight safety. As the EUROCONTROL CARE INO project SCOPE has demonstrated, increasing the reliability of pilot-controller communications using Automatic Speech Recognition (ASR) and advanced information presentation is feasible in a laboratory environment. In the SCOPE approach, the ASR system is used to track the dialogue between the controller and the pilot and the results of the tracking are presented to the controller by way of an enhanced interface to the radar display.

The ESCALE project is a follow-up to SCOPE in order to enable the scaling of the SCOPE approach from laboratory environment to real operational conditions, building on the solutions demonstrated in SCOPE and making the tracking system reliable in an operational context. Moreover, ESCALE proposes to identify innovative applications of the controller-pilot dialogue tracking paradigm, in collaboration with end-users, and aims at designing and implementing one of these applications in order to assess the usefulness of the tracking solution in an operational environment.

2. Objectives

The SCOPE project has proved the feasibility of the ASR-based dialogue tracking approach in laboratory environment, i.e., with a high bandwidth for the communication channel as
well as users with an average English accent and speaking rather slowly and clearly. This was sufficient for a demonstration of feasibility. Nevertheless, if one has in mind an operational use of the SCOPE solution, important issues specific to Air Traffic Control (ATC) communications, such as limited bandwidth and stressed speech, cannot be avoided, for these factors may generate an important amount of speech recognition errors. As a follow-up to SCOPE, the purpose of the ESCALE project is threefold:

- Enable the scaling of the SCOPE approach addressing the various issues pointed out as fundamental for an operational use of the controller-pilot dialogue tracking solution: reliability with limited bandwidth and stressed speech as well as integration into an operational environment;
- Propose a new approach to speech recognition error repair for the tracking solution, using rule-based consistence checking at the utterance level, relying upon domain-specific knowledge as well as dialogue context;
- In collaboration with experts from ATC and Avionics and operational users (controllers and pilots), identify innovative applications of the tracking functionality, such as redundancy of information for controllers/pilots in order to reduce cognitive load, and design and implement one of these applications as a prototype in order to demonstrate the reliability and usefulness of the controller-pilot dialogue tracking.

The first two objectives, as well as the identification of potential applications concern the first phase of the project. This phase is thus mainly concerned with the assessment of the scalability of the SCOPE approach to operational conditions. The second phase will be concerned mainly by the implementation of the selected application and its demonstration in an operational environment.

In order to address these objectives, the ESCALE consortium is composed of experts in ASR technology and applications as well as enhanced adaptive interfaces. The leader, THALES Research & Technology (TRT), has a strong know-how in ASR and language processing applications, in particular for the Air Traffic Management domain and the Avionics domain which are two of the major markets of the THALES group. The Institut de Recherche en Informatique de Toulouse (IRIT) features specialists in vocal and multimodal interaction, in particular for the ATC domain. IntuiLab is composed of experienced multimodal interface designers and developers which come from the ATC world. THALES Avionics (THAV) provides ESCALE with expertise concerning the integration into cockpits and the operational environment at the pilot side. And finally, the Technical University of Crete (TUC) features experts in acoustic modeling for ASR.

### 3. Operational ASR-based dialogue tracking

ATC communications have very specific characteristics, which are the following:\(^1\)

- Constrained domain-specific language (recurrent utterances, use of callsigns) and limited vocabulary;
- Important variability of speakers (potentially bad accents) with often no availability for training an ASR system;
- Noisy environment (cockpit, control rooms) and poor quality transmission channels (radio communications);
- Stressed speech: rapid delivery, bad pronunciation, interrupted or overlapping utterances.

In more technical terms, ASR for ATC communications faces the following constraints:\(^2\):

---


Limited bandwidth (300 to 3300 Hz) of VHF voice communications; English language has major energies in higher parts, i.e. “th” (8 to 9 kHz);

Technical instability at beginning of speech (VHF);

Incomplete words (phonemes) at the beginning, i.e., “fihansa one two” due to the use of “Push To Talk” switches;

Spontaneous speech with repetitions, hesitations: “mmh”, “haa” (approximately 4 to 5% of the utterances);

Mainly non native English speakers;

Mix of official ICAO languages (English, French, Spanish, Russian), i.e. “Geneva” in approximately 16% of the French utterances;

Pronunciation of navigation point names in the speaker’s mother tongue language;

Special keywords: break, disregard, correction.

In order to take these constraints into account for the design of a robust speech tracking system, the ESCALE approach consists in relying on a corpus of live recordings of communications between controllers in pilots. The project focuses on communications in English, for which only 13 hours are available, the remainder being either communications in French or blanks. The Vocalise corpus has been recorded in live conditions using standard recording facilities (speakers, microphone, recorder). Therefore, the recordings are very noisy with an important level of saturation. No information is available about the signal/noise ratio but many utterances are barely understandable or not understandable at all due to the saturation. Several other live recording specificities can been found in the corpus, such as incomplete words, in particular at the beginning of utterances, or breath.

The first task then consisted in isolating the usable data, that is the understandable data, and adapting it, which meant listening to all the utterances and writing down all the corresponding transcriptions, adjusting simultaneously the corpus segmentation. It appeared during this task that the Vocalise corpus is very difficult to process, even for a human trained to the operational phraseology. As such, the use of the SCOPE ASR system without the addition of specific acoustic models may not be possible.

It also appeared during this task that the modeling of the phraseology is very complicated. For instance, a specificity of the corpus is that French speakers tend to use French words in the middle of English sentences which increases the complexity of the phraseology. Due to this complexity, the use of a formal grammar to constrain the ASR engine may not be possible, as this grammar might be quite huge, mainly because of the variability in the use of the vocabulary of air navigation.

3.1. A corpus of live recordings

The corpus used in ESCALE is the Vocalise corpus from the Direction Générale de l’Aviation Civile (DGAC). It was the only European corpus easily available when the project started. It features 60 hours of live recordings of communications between French controllers and French/non-French pilots. The Vocalise corpus has been recorded in live conditions using standard recording facilities (speakers, microphone, recorder). Therefore, the recordings are very noisy with an important level of saturation. No information is available about the signal/noise ratio but many utterances are barely understandable or not understandable at all due to the saturation. Several other live recording specificities can be found in the corpus, such as incomplete words, in particular at the beginning of utterances, or breath.

The first task then consisted in isolating the usable data, that is the understandable data, and adapting it, which meant listening to all the utterances and writing down all the corresponding transcriptions, adjusting simultaneously the corpus segmentation. It appeared during this task that the Vocalise corpus is very difficult to process, even for a human trained to the operational phraseology. As such, the use of the SCOPE ASR system without the addition of specific acoustic models may not be possible.

It also appeared during this task that the modeling of the phraseology is very complicated. For instance, a specificity of the corpus is that French speakers tend to use French words in the middle of English sentences which increases the complexity of the phraseology. Due to this complexity, the use of a formal grammar to constrain the ASR engine may not be possible, as this grammar might be quite huge, mainly because of the variability in the use of the vocabulary of air navigation.

3.2. Ground reference for the tracking system

A first evaluation has been performed as to establish a ground reference for the tracking system. In particular, it was necessary to assess the possibility to use a formal grammar and the efficiency of the ASR engine using that grammar. The SCOPE context-free
grammar has been rebuilt in order to cover the phraseology from the Vocalise corpus. Tests have been performed on a subset of 600 utterances from the English part of the corpus in order to determine the efficiency of the speech recognition engine on the utterances parsed by the extended SCOPE grammar. These tests consider the efficiency from a semantic point of view. Thus, an utterance is considered as recognized not because all the words have been perfectly recognized but only if the command and the corresponding parameters values are those of the human transcription. For the controllers, few slots are recognized, where slots corresponds to a basic command component (either the callsign, an instruction or a parameter value). This result implies that recognized utterances are very different from the transcribed utterances. The only recognized slots correspond to instructions. All parameter values are wrong. Thus, in terms of semantic coherence, the recognized utterances do not retrieve at all the meaning of transcribed utterances. Nevertheless, the results show that almost 20% of utterances have at least one slot recognized. But, after a thorough study of the interpretations, it appears that a large part of recognized slots have been recognized “luckily”. In particular, the “proceed” instruction is often recognized because the speech recognition engine tends to produce series of VOR corresponding to that particular instruction. As an example, for the utterance “regional two seven one sierra charlie bonjour proceed direct arpus”, the semantic interpretation corresponding to the transcription is \{<callsign RG271SC> <instruction5 proceed> <value5 arpus>\}. One of the utterance produced by the speech recognition engine is “lima hotel terni motal vagna direct” and its semantic interpretation is \{<callsign PTMLH> <instruction5 proceed> <value5 terni motal vagna>\}. The recognized slot is <instruction5 proceed> and it is obvious that, despite the recognition of one common slot, the two utterances considered are far from each other from a semantic point of view. For the pilots, the percentage of recognized slots is also very low. The reason for these results is that the recognition stage is much more difficult because of the deterioration of the signal due to the saturation of the recordings. As very few recognized slots are relevant for both the controllers and the pilots, even with an appropriate grammar, and as instructions recognized without the corresponding parameter value have no semantic value and are therefore unusable, it appears that the use of specific acoustic models trained on the corpus becomes mandatory. Additional tests have been made on the transcriptions of the utterances from the subset of the corpus. About 50% of the utterances of the controllers are recognized by the extended SCOPE grammar, which is quite a low rate. For pilots, the results are better but the recognition rate is only about 65%. On the one hand, the better results obtained for pilots can be explained by the simplicity of the structures used by pilots and the fact that they rarely use reformulations, that is structures which cannot be modeled by a formal grammar. On the other hand, controllers tend to produce complex sentences. The variability between speakers is then more important and some utterance configurations cannot be modeled by a formal grammar. Despite these slight differences between controllers and pilots, it is clear from the results that the rates obtained in both cases are far from being satisfying and that such a grammar is not able to model the whole complexity of the phraseology. On the controllers’ side, utterances tend to be close to natural language, which cannot be modeled by a context-free grammar. On the pilots’ side, problems are due to the addition of short words anywhere in utterances. This variability is of huge complexity and cannot be modeled by a context-free grammar. The use of a stochastic language model trained on the corpus becomes thus mandatory.

### 3.3. Acoustic models

Acoustic models are in the process of being created relying upon training on the Vocalise corpus in order to adapt English phonemes to operational conditions and to replace French phonemes by the nearest English phonemes.
The variability between nationalities and accents of speakers makes the creation of robust acoustic models become a real challenge. This issue is rather difficult to address due to the overall audio quality of the recordings (noise, saturation, …) and the amount of training data, which is far from sufficient. However, the phoneme set that will be used for training has been defined. Additionally, the appropriate phoneme mapping from French to American English have been made. After a preliminary data preprocessing, a first baseline experiment has been run which showed that further data correction should be made in order to obtain a word error rate that does not approach 100%. This led to additional work on the corpus before the final acoustic modeling could be created.

3.4. Language model

The language model for ESCALE is also in the process of being constructed, relying upon training on the Vocalise corpus. A study of the disfluencies and corrections present in the corpus has been made, as such speech phenomena typically cannot be recognized with a formal grammar and thus require the use of a stochastic language model. Disfluencies and corrections in the Vocalise data include mainly:

- Hesitations;
- Repetitions;
- Pauses or breathing, inside utterances or inside the words themselves;
- Auto-corrections;
- Grammatical errors and other strange uses of language.

It appeared from this study that not all disfluencies and corrections were annotated in the corpus. Once again, this lead to an important amount of supplementary work to be performed on the corpus, with a complete re-listening and annotation of the 13 hours of usable data, before the final language model could be constructed.

4. Intelligent dialogue tracking

Intelligent dialogue tracking consists in using non audio-based ASR error detection and correction strategies coupled with the robust dialogue tracking system. These strategies are applied to the controller’s clearances and the pilot’s acknowledgement in order to detect and correct ASR errors which would remain after the ASR process. A set of basic rules is used: simple coherence rules such as for example that a plane cannot descend to a level which is higher than its actual level, or more domain-specific rules such as for example that a given model of airplane cannot perform the requested clearance. These rules are processed on the results of the tracking in order to check the validity of these results and detect additional errors and potentially repair them automatically or else alert the controller. More precisely, the rationale for the intelligent tracking approach is to build on approaches that rely upon the dialogue context and which study the cognitive processes likely to intervene in the communication failures, in particular the stress, the cognitive load and other characteristic of the live recordings. As such, error detection and correction uses contextual information, that is the history of the dialogue between the controller and a given pilot, and coherence of the dialogue is checked in order to detect the remaining errors.

5. Potential applications

The selection and the implementation of an application to serve as a benchmark for the ESCALE dialogue tracking functionality will be made in the second phase of the project, when the ASR-based tracking system will have been assessed. During the first phase, several candidate applications have been identified in collaboration with ATC and Avionics experts and end-users.
5.1. Attention getters

This application consists in extending the SCOPE application using visual or audio attention getters in cockpits. When the pilot’s dialog tracking module identifies in the flow of communications the callsign of his/her aircraft, it draws his/her attention using a visual or audio signal. The SCOPE demonstrator already used this kind of feedbacks in order to help controllers to link pilots which are speaking on the radio channel with the corresponding tracks on their radar display. The ESALE attention getters application is symmetric to the SCOPE one but for the cockpit. Such a feedback should make the pilot aware that he/she is the target of a specific message. Of course, this simple starter application can be coupled with more sophisticated functionalities for the pilot (see below). A particular attention will given to human factors while design this application, as feedbacks must inform the pilots without requiring too much of their cognitive and/or perceptual resources, which must be kept available for their main task.

5.2. Redundancy of information

This application consists in presenting visually to the pilot the controller’s clearance obtained from the ASR system, in addition to the standard audio channel. Since each communication act on the standard audio channel contains the callsign of the target aircraft, the ASR system is able to spot clearances and messages involving a particular aircraft in the controlled area. The information recognized by the ASR system can then be converted to a redundant modality, visual or textual for instance. Moreover, provided that the ATC communication language is well constrained and the vocabulary limited, informative clearances, that is other clearances than greetings and courtesy, can even be translated automatically into the mother’s tongue of the pilot. The feasibility of the translation approach has been demonstrated by TRT in the military domain which features operational languages with characteristics similar to the ATC communication language. However, a major human factors issue must be addressed with regards the translation of the clearance content. Indeed, presenting the messages to the pilot using a different language could be disturbing, making it more difficult to make the cognitive link between the oral message heard and the redundant presentation.

5.3. Clearance tracking

This application consists in extending the SCOPE application with new features:

- Checking that all the pilots make an acknowledgement in case of multiple intertwined clearances;
- Introducing a “clearance compliance” tracking system: whether operating a VFR or IFR flight, a pilot cannot deviate from a clearance without informing the controller in order to obtain an amended clearance. Whenever a clearance is acknowledged, the tracking system creates a “tracker” to check if the aircraft complies with the clearance without undue delay. In case of deviation, a feedback draws the controller’s attention to the aircraft track. For example, when the safety of an aircraft is jeopardized, pilots may deviate from clearances and may not notify controllers immediately. In such a case, the “clearance compliance” tracker displays a visual feedback until contact can be established;
- Taking into account communications initiated by the pilot: for instance, when a pilot announces himself as entering in a new sector, the corresponding track on the radar display is highlighted until the controller answers. If the controller does not, a timeout may stop the feedback, avoiding overloading the radar display with too much feedbacks;
On the aircraft side, displaying feedbacks after timeouts in order to recall the pilot that he/she did not acknowledge the last clearance or that the controller did not answer his/her last message. For instance, the feedback can be a display summarizing the last clearance intended to the aircraft. This kind of display can help the pilot to acknowledge the clearances more quickly and avoid un-acknowledged clearances.

5.4. Non-intrusive sensors

In this application, the dialogue tracking system is viewed as a tool to help human factors experts analyze controller interactions putting the controller in the interaction loop. Controllers naturally interact with other controllers and pilots, communicating and generating valuable information. Therefore, the most natural way to analyze controller interactions is to extract relevant information directly from observing these interactions. The tracking system is then used to develop non-intrusive sensors, for example to detect vocal emotion expressions such as happiness, anger, sadness, fear or normal state (i.e., “unemotional” state) or to detect auto-corrections, analyze the density and the nature of the controller dialogue with various pilots as a non-intrusive approach to detect stress or fatigue. The tracking system can also be coupled with a gaze tracker system in order to help disambiguating the audio information.

5.5. Automated transcription and annotation

Currently, it is widely impractical to search ATC audio archives by hand in order to find a recording containing a particular communication. In contrast, finding the text of a particular utterance in a textual archive can take only few minutes of search, using keywords. This application consists in transcribing and annotating audio data with a transcription of ATC communications to facilitate the work on large archives. With a text-like description attached to audio recording, the recording or a desired section can be easily found by searching on a description. The transcripts are in the form of time-stamped sentences prefixed by a callsign or the word “controller”. The discourse style of transcript is synthetic: it focuses on a lexicon of meaningful words obtained from the ICAO phraseology. Other types of annotation are also possible, such as the automatic transcription of prosody, emphasis and intonation, as well as speaker and dialect recognition. This enables searches based on expressive or emotional content, emphasis, speech forms such as questions, etc. Such annotations may help the BAE of the DGAC when investigating plane accidents. A controller can also need to review specific sections of recorded communications. A benefit of using a textual format to record relevant information is to enable the use of a search-engine to look for and play back specific records using a variety of search criteria (time, callsign, …). Since all the textual records are time-stamped, it is easy to make the link between a specific communicative act textual record and the corresponding section of the audio recordings. These textual recordings may also provide relief controllers with a new way to understand the current sector’s context when they take their services.

5.6. ATC simulation and pseudo-pilots

The cost of acquiring and running new and up-to-date ATC simulators is constantly escalating. These simulators tend to be resource intensive in terms of both space (i.e., a 360° airport tower simulator tends to employ a larger number of cubic meters) and manpower (instructors, pseudo-pilots, …). In the training field, ASR systems can be used to automate the task currently performed by pseudo-pilots. Pseudo-pilots are trained operators who manually enter data to drive simulators during training so that the simulated aircraft responds to the instructions received from a student controller. Being a pseudo-pilot role requires knowledge of ATC
phraseology and of simulator commands. Running an ATC exercise often requires several pseudo-pilots, thereby increasing the cost of using the simulator and sometimes skewing the exercise. Indeed, pseudo-pilot behavior can vary from “friendly” pseudo-pilots who may relax phraseology requirements and make the exercise easier to handle, to “perfectionist” pseudo-pilots who demand perfection in control techniques and phraseology before acting (acknowledgement, simulator inputs, …). Pseudo-pilot software using ASR systems and vocal synthesis can be used to provide several levels of automation:

- Analyzing the clearances of the training controller in order to fill a form with data to drive the simulator. The pseudo-pilot can then respond by an acknowledgement matching or not the clearance;
- Responding to a clearance according to a predefined scenario or rules such as “non-matching acknowledgements every five clearances”;
- Producing messages from aircrafts on the radio channel with a text-to-speech system with different voices, accents, gender, prosody, increasing the realism of the simulation;
- Producing an appropriate background noise to modulate pilot responses in order to increase realism;
- Enlarging the scope of pseudo-pilots: an intelligent pseudo-pilot may not only react to the student’s voice and respond to his/her commands, but also observe the student’s performance and offer advices, and initiate the dialogue. For instance, the simulated pilot can ask to fly at the optimal flight level according to the aircraft’s specifications.

Such software might reduce the number of human pseudo-pilots needed in order to run the ATC exercises, reducing the cost of such exercises while increasing their realism. It is even possible to create a fully automated pseudo-pilot following a predefined scenario when reacting to the training controller’s clearances. Moreover, automated pseudo-pilots can be used in an open system, enabling to connect other agents according to the needs of the simulation. For example, during an exercise, it is possible to simulate an aircraft using a flight simulator permitting to train a pilot student to the air traffic conditions.

6. Conclusion

The objective of the ESCALE project is to increase the reliability of controller-pilot communications and thus increase the overall ATC safety, relying upon a tracking of the dialogue between the controller and the pilot, in order to detect misunderstandings. The first phase of the ESCALE project concerned the assessment of the scalability of the SCOPE dialogue tracking approach to an operational environment. From tests performed on a corpus of live recordings of controller-pilot communications, namely the Vocalise corpus from the DGAC, it appears that the SCOPE ASR system, which is based on a formal grammar and a non specific recognition engine, does not provide sufficient performances under operational conditions. The use of a stochastic language model and of specific acoustic models becomes then mandatory. The creation of these models, relying upon training on the Vocalise data, is still in progress.

Several candidate applications of the dialogue tracking functionality have been identified in collaboration with experts and end-users from the ATC and Avionics worlds. The selection and implementation of a given application, in order to demonstrate the ESCALE approach in an operational environment, will be done in the second phase of the project.