The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech

Konrad Hofbauer[*], **Stefan Petrik**[*], **Horst Hering**[†]

[*]Graz University of Technology
Signal Processing and Speech Communication Laboratory, Inffeldgasse 12, 8010 Graz, Austria
{konrad.hofbauer, stefan.petrik}@tugraz.at
[†] EUROCONTROL Experimental Centre
Centre du Bois des Bordes B.P. 15, 91222 Brétigny-sur-Orge CEDEX, France
horst.hering@eurocontrol.int

Abstract

Air traffic control (ATC) is based on voice communication between pilots and controllers and uses a highly task and domain specific language. Due to this very reason, spoken language technologies for ATC require domain-specific corpora, of which only few exist to this day. The ATCOSIM Air Traffic Control Simulation Speech corpus is a speech database of non-prompted and clean ATC operator speech. It consists of ten hours of speech data, which were recorded in typical ATC control room conditions during ATC real-time simulations. The database includes orthographic transcriptions and additional information on speakers and recording sessions. The ATCOSIM corpus is publicly available and provided online free of charge. In this paper, we first give an overview of ATC related corpora and their shortcomings. We then show the difficulties in obtaining operational ATC speech recordings and propose the use of existing ATC real-time simulations. We describe the recording, transcription, production and validation process of the ATCOSIM corpus, and outline an application example for automatic speech recognition in the ATC domain.

1. Introduction

The corpus presented in this document is within the domain of civil *air traffic control (ATC)*. The aim of air traffic control is to maintain a safe separation between all aircraft in the air in order to avoid collisions, and to maximise the number of aircraft that can fly at the same time. Besides a set of fixed flight rules and a number of navigational systems, air traffic control relies on human air traffic control operators (ATCO, or *controllers*). The controller monitors air traffic within a so-called *sector* (a geographic region or airspace volume) based on previously submitted flight plans and continuously updated radar pictures, and gives flight instructions to the aircraft pilots in order to maintain a safe separation between the aircraft.

Although digital data communication links between controllers and aircraft are slowly emerging, most of the communication between controllers and pilots is verbal and by means of analogue voice radios. The communication occurs on a party-line channel, which means that all aircraft within a sector as well as the corresponding controller can hear all messages that are transmitted on that radio channel frequency.

The international standard language for ATC communication is English. The use of French, Spanish or Russian language is also permitted if it is the native language of both pilot and controller involved in the communication. The phraseology that is used for this communication is strictly formalised by the International Civil Aviation Organization (ICAO, 2006). It mandates the use of certain keywords and expressions for certain types of instructions, gives clear rules on how to form digit sequences, and even defines non-standard pronunciations for certain words in order to account for the band-limited transmission channel. In practise however, both controllers and pilots deviate from this standard phraseology.

Until today spoken language technologies such as automatic speech recognition are close to non-existent in operational air traffic control. This is in parts due to the high reliability requirements that are naturally present in air traffic control. The constant progress in the development of spoken language technologies more and more opens a door to the use of such techniques for certain applications in the air traffic control domain. This is particularly the case for the controller speech on ground, considering the good signal quality (close-talk microphone, low background noise, known speaker) and the restricted vocabulary and grammar in use. In contrast, doing for example speech recognition for the incoming noisy and narrowband radio speech is still a quite difficult task.

In the development of practical systems the need for appropriate corpora comes into place. The quality of air traffic control speech is quite particular and falls in-between the classical categories: It is neither spontaneous speech due to the given constraints, nor is it read, nor is it a pure command and control speech (in the sense of controlling a device). Due to this and also due to the particular pronunciation and vocabulary in air traffic control, there is a need for speech corpora that are specific to air traffic control. This is even more the case considering the high accuracy and robustness requirements in most air traffic control applications.

We review in Section 2. the few existing ATC related corpora known to the authors. The subsequent sections present the new ATCOSIM Air Traffic Control Simulation Speech corpus, which fills a gap that is left by the existing corpora. Section 3. outlines the difficulty of obtaining realistic air traffic control speech recordings and shows the path chosen for the ATCOSIM corpus. The transcription and production process is described in Section 4. and 5., whereas Section 6. presents the validation process chosen for the ATCOSIM corpus. We conclude with a proposal for a specific ASR application in air traffic control.

2. ATC Related Corpora

Despite the large number of existing corpora, only a few corpora are in the air traffic control domain.

The *NIST Air Traffic Control Complete Corpus* (Godfrey, 1994) consists of recordings of 70 hours of approach control radio transmissions at three airports in the United States. The recordings are narrowband and of typical AM radio quality. The corpus contains an orthographic transcription and for each transmission the corresponding flight number is listed. The corpus was produced in 1994 and is commercially available.

The *HIWIRE* database (Segura et al., 2007) is a collection of read or prompted words and sentences taken from the area of military air traffic control. The recordings were made in a studio setting, and cockpit noise was artificially added afterwards. The database contains 8,100 English utterances pronounced by non-native speakers without air traffic control experience. The corpus is available on request.

The *non-native Military Air Traffic Control (nnMATC)* database (Pigeon et al., 2007) is a collection of 24 hours of military ATC radio speech. The recordings were made in a military air traffic control centre, wire-tapping the actual radio communication during military exercises. The recordings are narrowband and of varying quality depending on the speaker location (control room or aircraft). The database was published in 2007, but its use is restricted to the NATO/RTO/IST-031 working group and its affiliates.

The *VOCALISE* project (Graglia et al., 2005; Arnoux et al., 2005) recorded and analysed 150 hours of operational ATC voice radio communication in France, including en route, approach and tower control. The database is not available for the public and its use is restricted to research groups affiliated with the French 'Centre d'Études de la Navigation Aérienne' (CENA)—now part of the 'Direction des Services de la Navigation Aérienne' (DSNA).

The aforementioned corpora vary significantly among each other with respect to e.g. scope, technical conditions or public availability (Table 1). The aim of the ATCOSIM corpus is to fill the gap that is left by the above corpora: ATCOSIM provides 50 hours of publicly available direct-microphone recordings of operational air traffic controller speech in a realistic civil en-route control situation. The corpus includes an utterance segmentation and an orthographic transcription. ATCOSIM is meant to be versatile and is as such not tailored to any specific application.

3. ATCOSIM Recording and Processing

The aim of the ATCOSIM corpus production was to provide wideband ATC speech which should be as realistic as possible in terms of speaking style, language use, background noise, stress levels, etc.

In most air traffic control centres the controller pilot radio communication is recorded and archived for legal reasons. However, these legal recordings are problematic for a corpus production for a multitude of reasons. First, most recordings are based on a received radio signal and thus not wideband. Second, it is in general difficult to get access to these recordings. And third, even if one would obtain the recordings, their public distribution would be legally problematic at least in many European countries.

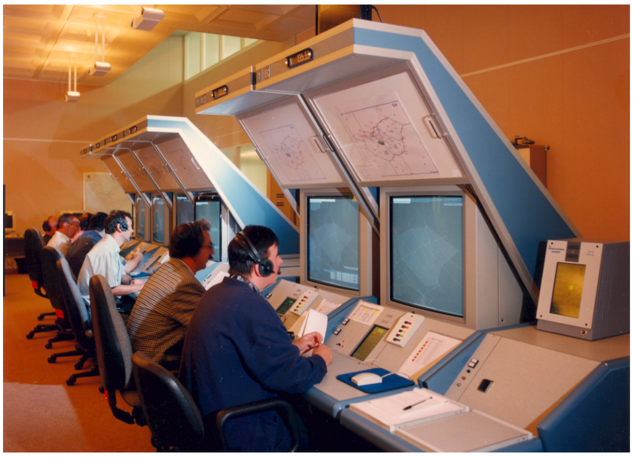


Figure 1: Control room and controller working position at the EUROCONTROL Experimental Centre (recording site)

The logical resort is the conduction of simulations in order to generate the speech samples. However, the required effort to set up realistic ATC simulations (including facilities, hard- and software, trained personal, . . .) is large and would well exceed the budget of a typical corpora production. However, such simulations are performed for the sake of evaluating air traffic control and air traffic management concepts, also on a large scale involving tens of controllers. The ATCOSIM speech recordings were made at such a facility during an ongoing simulation.

3.1. Recording Situation

The voice recordings were made in the air traffic control room of the EUROCONTROL Experimental Centre (EEC) in Brétigny-sur-Orge, France (Figure 1). The room and its controller working positions closely resemble an operational control centre room. The simulations aim to provide realistic air traffic scenarios and working conditions for the air traffic controller. Several controllers operate at the same time, in order to simulate also the inter-dependencies between different control sectors. The controller communicates via a headset with pseudo-pilots which are located in a different room and control the simulated aircraft. During the simulations only the controllers' voice, but not the pilots', was recorded, because the working environment of the pseudo-pilots, and as such the speaking style, did not to any extent resemble reality.

3.2. Speakers

The participating controllers were all actively employed air traffic controllers and possessed professional experience in the simulated sectors. The six male and four female controllers were of either German or Swiss nationality and had German, Swiss German or Swiss French native tongue. The controllers had agreed to the recording of their voice for the purpose of language analysis as well as for research and development in speech technologies, and were asked to show their normal working behaviour.

3.3. Recording Setup

The controller's speech was picked up by the microphone of a Sennheiser HME 45-KA headset. The microphone signal

	NIST	**HIWIRE**	**nnMATC**	**VOCALISE**	**ATCOSIM**
Recording Situation					
- Recording content	civil ATCO & PLT	N/A	military ATCO & PLT	civil ATCO & PLT	**civil ATCO**
- Control position	approach	N/A	military	mixed	**en-route**
- Geographic region	USA	N/A	Europe (BE)	Europe (FR)	**Europe (DE/CH/FR)**
- Speaking style (context)	operational	prompted text	operational	operational	**operational** [1]
Recording Setup					
- Speech bandwidth	narrowband	wideband	mostly narrowband	unknown	**wideband**
- Transmission channel	radio	none	none / radio	none / radio	**none**
- Radio transmission noise	high	none	mixed	mixed	**none**
- Acoustical noise	CO & CR	CO (artificial)	CO & CR	CO & CR	**CR**
- Signal source	VHF radio	direct microphone	mixed	unknown	**direct microphone**
Speaker Properties					
- Level of English	mostly native	non-native	mostly non-native	mixed	**non-native**
- Gender	mixed	mixed	mostly male	mixed	**mixed**
- Operational	yes	no (!)	yes	yes	**yes**
- Field of prof. operation	civil	N/A	military	civil	**civil**
- Number of speakers	unknown (large)	81	unknown (large)	unknown (large)	**10**
Publicly Available	yes	yes (?)	no	no	**yes**

[1] Large-scale real-time simulation

• CO: Cockpit • CR: Control Room • ATCO: Controller • PLT: Pilot

Table 1: Feature comparison of existing ATC-related speech corpora and the ATCOSIM corpus.

and a push-to-talk (PTT) switch status signal were recorded onto digital audio tape (DAT) with a sampling frequency of 32 kHz and a resolution of 12 bit. The push-to-talk switch is the push-button that the controller has to press and hold in order to transmit the voice signal on the real-world radio. The speech signal was automatically muted when the push-button was not pressed. This results in a truncation of the speech signal if the push-button was pressed too late or released too early. Figure 2 shows an example of the recorded signals. After the digital transfer of the DAT tapes onto a personal computer, the status signal of the push-to-talk button could after some basic processing be used to reliably perform an automatic segmentation of the recorded voice signal into separate controller utterances.

4. Orthographic Transcription

The speech corpus includes an orthographic transcription of the controller utterances. The orthographic transcriptions are aligned with each utterance.

4.1. Transcription Environment

The open-source tool TableTrans was chosen for the transcription of the corpus (Maeda et al., 2002). TableTrans was selected for its table-based input structure as well as for its capability to readily import the automatic segmentation. The transcriptionist fills out a table in the upper half of the window where each row represents one utterance. In the lower half of the window the waveform of the utterance that is currently selected or edited in the table is automatically displayed (Figure 3). The transcriptionist can play, pause and replay the currently active utterance by a single key stroke or as well select and play a certain segment in the waveform display. A small number of minor modifications to the TableTrans applications were made in order to lock certain user interface elements and to extend its replay capabilities.

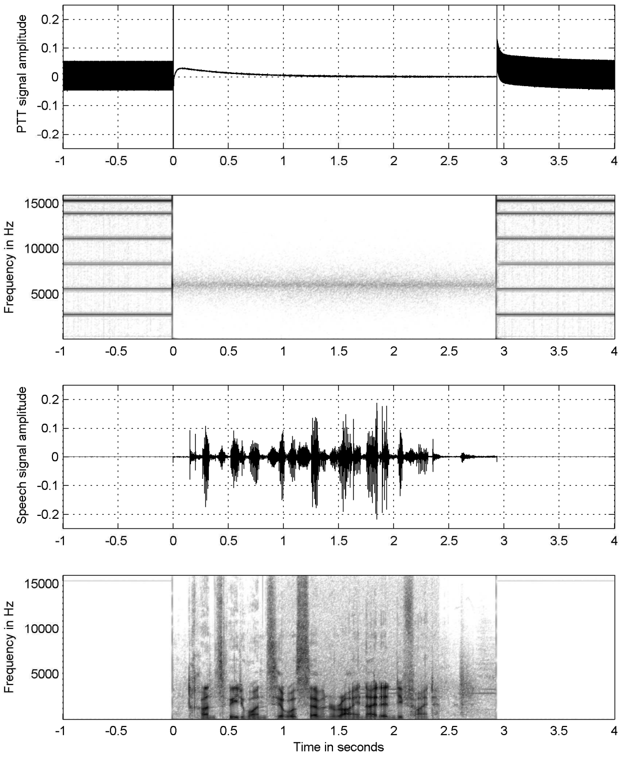
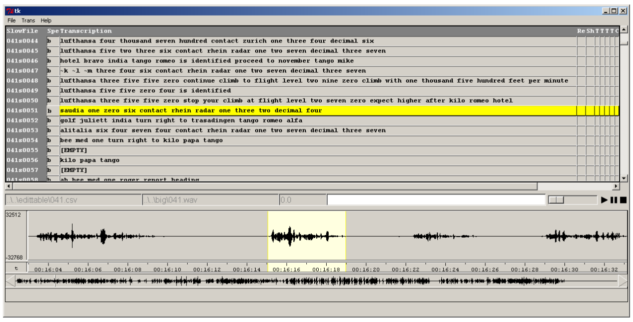


Figure 2: A short speech segment (`transwede one zero seven rhein identified`) with push-to-talk (PTT) signal. Time domain signal and spectrogram of the PTT signal (top two) and time-domain signal and spectrogram of the speech signal (bottom two)

A number of keyboard shortcuts were provided to the transcriptionist using the open-source tool AutoHotKey (Mallett, 2008). These were used for conveniently accessing alter-



SlowFile	Spe	Transcription	Re	Sh	T	T	T	T	C
041s0044	b	lufthansa four thousand seven hundred contact zurich one three four decimal six							
041s0045	b	lufthansa five two three six contact rhein radar one two seven decimal three seven							
041s0046	b	hotel bravo india tango romeo is identified proceed to november tango mike							
041s0047	b	~k ~l ~m three four six contact rhein radar one two seven decimal three seven							
041s0048	b	lufthansa three five five zero continue climb to flight level two nine zero climb with one thousand five hundred feet per minute							
041s0049	b	lufthansa five five zero four is identified							
041s0050	b	lufthansa three five five zero stop your climb at flight level two seven zero expect higher after kilo romeo hotel							
041s0051	b	saudia one zero six contact rhein radar one three two decimal four							
041s0052	b	golf juliett india turn right to trasadingen tango romeo alfa							
041s0053	b	alitalia six four seven four contact rhein radar one two seven decimal three seven							
041s0054	b	bee med one turn right to kilo papa tango							
041s0055	b	[EMPTY]							
041s0056	b	kilo papa tango							
041s0057	b	[EMPTY]							
041s0058	b	ah bee med one roger report heading							

Figure 3: Screen-shot of the transcription tool TableTrans

native time-stretched sound files[1] and for entering frequent character or word sequences such as pre-defined keywords, ICAO alphabet spellings and frequent commands, both for convenience and in order to avoid typing mistakes.

4.2. Transcription Format

The orthographic transcription follows a strict set of rules which is included in the corpus documentation. In general, all utterances are transcribed word-for-word in standard British English. All standard text is written in lower-case. Punctuation marks including periods, commas and hyphens are omitted. Apostrophes are used only for possessives (e.g. `pilot's radio`)[2] and for standard English contractions (e.g. `it's`, `don't`). Numbers, letters, navigational aids and radio call signs are transcribed following a given definition based on several aeronautical standards and references. Regular letters and words are preceded or followed by special characters to mark truncations (=), individually pronounced letters (~) or unconfirmed airline names (@).

Stand-alone technical mark-up and meta tags are written in upper case letters with enclosing squared brackets. They denote human noises such as coughing, laughing and sighs (`[HNOISE]`), fragments of words (`[FRAGMENT]`), empty utterances (`[EMPTY]`), nonsensical words (`[NONSENSE]`), and unknown words (`[UNKNOWN]`). Groups of words are embraced by opening and closing XML-style tags to mark off-talk (`<OT> ... </OT>`), which is also transcribed, and foreign language (`<FL> </FL>`), for which a transcription could be added at a later stage. Table 2 gives several examples of transcribed utterances.

Silent pauses both between and within words are not transcribed. For consistent results this would require an objective measure and criterion and is thus easier to integrate in combination with a potential future word segmentation of the corpus. Also technical noises as well as speech and noises in the background—produced by speakers other than the one recorded—are not transcribed, as they are virtually always present and are part of the typical acoustical situation in an air traffic control room.

aero lloyd five one seven proceed direct to frankfurt
speedway three three five two contact milan one three four five two bye
ah lufthansa five five zero four turn right ten degrees report new heading
hapag lloyd six five three in radar contact climb to level two nine zero
scandinavian six one seven proceed to fribourg fox romeo india
ind= israeli air force six eight six resume on navigation to tango
good afternoon belgian airforce forty four non ~r ~v ~s ~m identified
[HNOISE] alitalia six four seven four report your heading
[FRAGMENT] hapag lloyd one one two identified
sata nine six zero one is identified <OT> oh it's over now </OT>
aero lloyd [UNKNOWN] charlie papa alfa guten tag radar contact

Table 2: Examples of controller utterance transcriptions in the ATCOSIM corpus
Human noises are labelled with `[HNOISE]`, word fragments with `[FRAGMENT]` and unintelligible words with `[UNKNOWN]`. Truncations are marked with an equals sign (=), individually pronounced letters with a tilde (~), and beginning and end of off-talk is labelled with `<OT>` and `</OT>`.

4.3. Transcription Process and Quality Assurance

The entire corpus was transcribed by a single person, which promises high consistency of the transcription across the entire database. The native English speaker was introduced to the basic ATC phraseology (ICAO, 2006) and given lists covering country-specific toponyms and radio call signs (e.g. (ICAO, 1994)). Clear transcription guidelines were established and new cases that were not yet covered by the guidelines immediately discussed.

Roughly three percent of all utterances were randomly selected across all speakers and used for a pre-training of the transcriptionist. This pre-transcription was also used to validate the applicability of the transcription format definition and minor changes were made. The transcriptions collected during the training phase were discarded and the material re-transcribed in the course of the final transcription.

After the transcription was finished, the transcriptionist once again reviewed all utterances, verified the transcriptions and applied corrections where necessary. Remaining unclear cases were shown to an operational air traffic controller and most of them resolved.

Due to the frequent occurrence of special location names and radio call signs an automatic spell check was not performed. Instead of this, a lexicon of all occurring words was created, which includes a count of occurrence and examples of the context in which the word occurs. Due to the limited vocabulary used in ATC, this list consists of less than one thousand entries including location names, call signs, truncated words, and special mark-up codes. Every item of the list was manually checked and thus typing errors eliminated.

5. ATCOSIM Structure and Distribution

The entire corpus including the recordings and all meta data has a size of approximately 2.5 gigabyte and is available in digital form on a single DVD or an electronic ISO disk image. Some statistics of the corpus are given in Table 3.

[1] The duration of each utterance was stretched by a factor of 1.7 using the PRAAT implementation of the PSOLA method (Boersma and Weenink, 2007). They were used only when dealing with utterances that were difficult to understand.

[2] Corpus transcription excerpts are written in a `mono-spaced typewriter` font.

Duration total (thereof speech)	51.4 h (10.7 h)
Data size	2.4 GB
Speakers (thereof female/male)	10 (4/6)
Sessions total	50
Sessions per speaker	7, 9, 5, 6, 1, 2, 2, 8, 7, 3
Utterances total	10078
Utterances per speaker	1167, 1848, 808, 1162, 238, 384, 378, 1716, 1739, 638
Utterance duration (mean, std. deviation, min, max)	3.8 s, 1.6 s, 0.04 s, 38.9 s
Utterances, containing	
- <FL> </FL>	182
- <OT> </OT>	84
- [EMPTY]	319
- [FRAGMENT]	35
- [HNOISE]	62
- [NONSENSE]	11
- [UNKNOWN]	11
Words	108883
Characters (thereof without space)	626425 (517542)
Lexicon entries	
- Total	858
- Meta tags	9
- Truncations	106
- Compounds	13
- Unique words	730

Table 3: Key figures of the ATCOSIM corpus

5.1. Corpus Structure and Format

The ATCOSIM corpus data is composed of four directories. The 'WAVdata' directory contains the recorded speech signal data as single-channel Microsoft WAVE files with a sample rate of 32 kHz and a resolution of 16 bits per sample. Each file corresponds to one controller utterance. The 10,078 files are located in a sub-directory structure with a separate directory for each of the ten speakers and subdirectories thereof for each simulation session of the speaker. The 'TXTdata' directory contains single text files with the orthographic transcription for each utterance. They are organised in the same way as the audio files. The directory also contains an alphabetically sorted lexicon and a comma-separated-value file which includes not only the transcription of all utterances but also all meta data such as speaker, utterance and session IDs and transcriptionist comments.
The files in the 'HTMLdata' directory are HTML files which present the transcriptions and the meta data in a table-like format. They enable immediate sorting, reviewing and replaying of utterances and transcriptions from within a standard HTML web browser. Last, the 'DOC' directory contains all documentation related to the corpus.

5.2. License and Distribution

The ATCOSIM corpus is available online at http://www.spsc.tugraz.at/ATCOSIM and provided free of charge. It can be freely used for research and development, also in a commercial environment. The corpus is also foreseen to be distributed on DVD through the European Language Resources Association (ELRA).

6. Validation

The validation of the database was carried out by the Signal Processing and Speech Communication Laboratory (SPSC) of Graz University of Technology, Austria. The examiner and author of the validation report has not been involved in the production of the ATCOSIM corpus, but only carried out an informal pre-validation and the formal final validation of the corpus.
The validation procedure followed the guidelines of the Bavarian Archive for Speech Signals (BAS) (Schiel and Draxler, 2003). It included a number of automatic tests concerning completeness, readability, and parsability of data, which were successfully performed without revealing errors. Furthermore, manual inspections of documentation, meta data, transcriptions, and the lexicon were done, which showed minor shortcomings that were fixed before the public release of the corpus. Finally, a re-transcription of 1% of the corpus data was made, showing a transcription accuracy on word level of 99.4%, proving the transcriptions to be accurate.
The ATCOSIM corpus was therefore considered to be in a usable state for speech technology applications.

7. Conclusion and Outlook

The ATCOSIM corpus is a valuable contribution to application-specific language resources. To our best knowledge currently no other speech corpus is publicly available that contains non-prompted air traffic control speech with direct microphone recordings, as it is difficult to produce such recordings for public distribution. The large-scale real-time ATC simulations exploited for this corpus production provide an opportunity to record ATC speech which is very similar to operational speech, while avoiding the legal hassle of recording operational ATC speech.
The application possibilities for spoken language technologies in air traffic control are manifold and we conclude with one concrete example: The controller sees on the radar screen in the text label corresponding to the aircraft among other information the current flight level of the aircraft. For example, a controller issues an instruction to an aircraft to climb from its current flight level 300 to flight level 340 (e.g. "sabena nine seven zero climb to flight level three four zero"). In certain ATC display systems, the controller now enters this information ('climb to 340') into the system and it shows up in the aircraft label, as this information is relevant later on when routing other adjacent aircraft. However, the voice radio message sent to the pilot already contains all information required by the system, namely the aircraft call-sign and the instruction.
Depending on the achievable robustness, an automatic speech recognition (ASR) system that recognises the controller's voice radio messages could perform various tasks: In case of extremely high accuracy the system could gather the information directly from the voice message without any user interaction. The ASR system could otherwise provide a small list of suggestions, to ease the process of entering the instructions into the system. Alternatively, the system could compare in the background the voice messages sent

to the pilot and the instructions entered into the system, and give a warning in case of discrepancies.

Compared to other ASR applications, the conditions would be comparably favourable in this scenario: The signal quality is high due to the use of a close-talk microphone and the absence of a transmission channel. The vocabulary is limited, and additional side information, such as the aircraft present in the sector and context-related constraints, can be exploited. The ASR system can be speaker-dependent and pre-trained, and continue training due do the constant feedback given by the controller during operation.

8. Acknowledgements

The authors are grateful for the the cooperation of the involved air traffic controllers, and want to acknowledge the fundamental contribution of the transcriptionist. The authors also wish to thank Gernot Kubin from Graz University of Technology for the helpful advice. The corpus was funded by the EUROCONTROL Experimental Centre, France.

9. References

Cyril Arnoux, Louis Graglia, and Didier Pavet. 2005. VOCALISE - the today use of VHF as a media for pilots/controllers communications. Online. http://www.cena.aviation-civile.gouv.fr/divisions/ICS/projets/vocalise/index_en.html.

Paul Boersma and David Weenink. 2007. PRAAT: doing phonetics by computer. Computer program. http://www.praat.org/.

John J. Godfrey. 1994. *Air Traffic Control Complete*. Linguistic Data Consortium, Philadelphia, U.S.A. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S14A.

L. Graglia, B. Favennec, and A. Arnoux. 2005. Vocalise: Assessing the impact of data link technology on the R/T channel. In *24th Digital Avionics Systems Conference (DASC)*, Washington D.C., U.S.A.

ICAO. 1994. *Designators for Aircraft Operating Agencies, Aeronautical Authorities and Services*. Number 8585/93. International Civil Aviation Organization.

ICAO. 2006. *Manual of Radiotelephony*. Number AN/925 in Doc 9432. International Civil Aviation Organization, 3 edition.

Kazuaki Maeda, Steven Bird, Xiaoyi Ma, and Haejoong Lee. 2002. Creating annotation tools with the annotation graph toolkit. In *Third International Conference on Language Resources and Evaluation (LREC)*, Paris, France.

Chris Mallett. 2008. Autohotkey - free mouse and keyboard macro program with hotkeys and autotext. Computer program. http://www.autohotkey.com/.

Stephane Pigeon, Wade Shen, and David van Leeuwen. 2007. Design and characterization of the non-native military air traffic communications database (nnMATC). In *International Conference on Spoken Language Processing (INTERSPEECH)*, Antwerp, Belgium.

Florian Schiel and Christoph Draxler. 2003. *Production and Validation of Speech Corpora*. Bastard Verlag München.

J.C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos. 2007. The HIWIRE database, a noisy and non-native english speech corpus for cockpit communication. Online. http://www.hiwire.org/.