A Real-life, French-accented Corpus of Air Traffic Control Communications

Estelle Delpech[1], Marion Laignelet[2], Christophe Pimm[2], Céline Raynal[2], Michal Trzos[3], Alexandre Arnold[1], Dominique Pronto[1]

[1]AIRBUS, Toulouse, France - [2]Safety Data-CFH, Toulouse, France - [3]HONEYWELL, Brno, Czech Republic

{estelle.e.delpech;alexandre.arnold;dominique.pronto}@airbus.com, {laignelet;pimm;raynal}@safety-data.com, michal.trzos@honeywell.com

Abstract

This paper describes the creation of the AIRBUS-ATC corpus, which is a real-life, French-accented speech corpus of Air Traffic Control (ATC) communications (message exchanged between pilots and controllers) intended to build a robust ATC speech recognition engine. The corpus is currently composed of 59 hours of transcribed English audio, along with linguistic and meta-data annotations. It is intended to reach 100 hours by the end of the project. We describe ATC speech specificities, how the audio is collected, transcribed and what techniques were used to ensure transcription quality while limiting transcription costs. A detailed description of the corpus content (speaker gender, accent, role, type of control, speech turn duration) is given. Finally, preliminary results obtained with state-of-the-art speech recognition techniques support the idea that accent-specific corpora will play a pivotal role in building robust ATC speech recognition applications.

Keywords: speech corpus, spoken language, controlled language, air traffic control phraseology, speech recognition, accented English.

1. Introduction

The AIRBUS-ATC corpus was developed in order to build a speech recognition system able to process Air Traffic Control (ATC) communications - messages exchanged between pilots and controllers using an English-based controlled language known as the ICAO phraseology (ICAO, 2007) - and Automatic Terminal Information Service messages (ATIS, airport information broadcast about weather conditions, available runways, etc.). The goal, in the long term, is to help pilots by giving them reliable ATC information in a persistent and visual way.

Pilot: Montana, F-CD, request cancel my IFR flight, proceeding VFR estimating Borton at 1701

Controller: F-CD, IFR flight cancelled at 35, contact Montanan Information 125.750

Figure 1: Example of ICAO phraseology.

ATC communications being very different from everyday conversations (see table 1), voluminous datasets like the SWITCHBOARD (Godfrey et al. 1992) and FISHER (Cieri et al. 2004) corpora cannot be used to build such a system although they were key in achieving human parity for conversational speech (Hannun et al. 2014; Xiong et al. 2016; Saon et al. 2017).

	SWB/FISHER	**ATC speech**
intelligibility	good (phone)	bad (radio transmission)
accents	US English	diverse and non-native
lexicon & syntax	oral syntax everyday topics	limited to ICAO phraseology or related
speech rate	standard	high (Cauldwell, 2007)
other	-	code switching

Table 1: SWB & FISHER corpora vs. ATC speech

The AIRBUS-ATC corpus intends to address this issue by providing a real-life corpus of transcribed English ATC messages spoken by non-native speakers.

2. Existing ATC speech corpora

Six ATC speech corpora were found in the literature. Three of them are unavailable: the nnMTAC corpus (Pigeon et al. 2007) – 24 hours of real-life, non-native military ATC, the VOCALISE dataset (Graglia et al. 2005) and the corpus of Lopez et al. (2013) – respectively 150h and 22h of real-life French-accented civil ATC.

Available resources are the ATCOSIM (Hofbauer et al. 2008), HIWIRE (Segura et al. 2007) and NIST (Godfrey; 1994) corpora. ATCOSIM and HIWIRE are rather small corpora (resp. 10,7 hours and 8,100 utterances of 1 to 12 words) containing various non-native accents (resp. 3 and 4 distinct accents). Their main limitation is that they were collected in simulated situations. ATCOSIM contains only controller messages uttered during training sessions and there is no radio transmission noise. HIWIRE is geared towards vocal commands and thus includes some datalink commands (pilot-controller texting tool limited to a subset of ATC messages). The text of the commands was generated with a deterministic grammar and then read. Cockpit noise was added afterwards and there is no radio transmission noise. The NIST Air Traffic Control Corpus (Godfrey, J., 1994) is the best fit for our goal: 70 hrs of real-life ATC from 3 different US airports. Still, it shall be extended with non-native data, which we expect to do with the AIRBUS-ATC corpus.

3. Corpus description

The corpus is composed of 2,160 paired audio + transcription files amounting to nearly 59 hrs of transcribed English. Audio files are mono-channel *.wav* files, with 16 kHz sampling rate and 32 bits resolution. Transcription files are in XML-based *.trs* format which is

the format of the transcription software *Transcriber*[1]. In addition to text and time stamps, the *.trs* format allows the encoding of information about speakers, recording context, speech turns, phonetic events and semantic entities.

3.1 Corpus characteristics

The corpus contains three types of Air Traffic Control: approach (APPR) and tower (TWR) controls, both coming from the same airport. It also contains ATIS recordings from 35 French airports. The audio contains both French and English speech, but only the English parts are transcribed.

Results in table 2 highlight that a huge quantity of raw audio is required to obtain English utterances alone, especially when the audio comes from French airports, where French and English are official ATC languages. As for ATIS, the same message is repeated until a change occurs in airport conditions which makes quite small the percentage of relevant data.3.2 Speech turn characteristics

APPR-TWR and ATIS speech turns have distinctive durations. ATIS turns are 6 times longer than APPR-TWR ones.

	Total duration	**English**	**% of English**
APPR-TWR	101:50:08	48:23:13	47.5%
ATIS w/ repetition	285:08:41	1:46:32	0.6%
ATIS w/o repetition	17:18:34	8:47:30	50.8%
TOTAL	404:17:23	**58:57:15**	**14.6%**

Table 2: Corpus content

	average turn duration (sec.)	**average of turn with foreign inclusion (%)**
APPR-TWR	4.4 sec	16.14%
ATIS	29 sec.	0.06%
TOTAL	**5.2 sec**	**15.65%**

Table 3: Speech turns characteristics

Around 16% of English APPR-TWR turns include at least one foreign word. Foreign words are forbidden by ICAO phraseology, these short occurrences correspond to courtesy words, ex. *bonjour* ("hello"), *merci* ("thank you").

3.3 Speakers and language characteristics

18 different native accents are indexed in the APPR-TWR part of the corpus. As expected, French accent is the most represented; other most frequent accents are English, German and Spanish. ATIS is once again distinctive because only spoken by local French natives.

Concerning the distribution of turns between controllers and pilots, speaking times between pilots and controllers are well-balanced as shown in Table 4. This is compliant with ATC phraseology rules: to one sentence emitted by the controller, an assessment from pilots shall follow with a repetition of the initial content that assesses the given instruction.

	% Pilots	**% Controllers**	**% not recog.**
APPR-TWR	55%	44.7%	0.3%

Table 4: Representativeness of roles

The ratio between men and women is disproportionate but it is representative of the real-life working situation.

	% Male	**% Female**	**% unknown**
APPR-TWR	75.3%	24.4%	0.3%
ATIS	68%	31%	1%
TOTAL	75%	24.6%	0.4%

Table 5: Gender representativeness

4. Corpus acquisition and processing

4.1 Audio collection

One key requirement is to collect ATC communications with audio quality as close as possible to real-life conditions. The chosen technical solution was to use a software-defined radio receiver connected to an aeronautical antenna and set to capture local airport APPR-TWR and ATIS broadcasts (~85% of corpus). This setup can collect up to 283 GB of audio data over a 30 days period. The remaining 15% were collected by automatically calling airport dedicated ATIS phone numbers. This means of collection is less tedious but does not provide audio with VHF quality.

4.2 Preprocessing

The raw audio files contain long sequences of silences and are too big to be processed by transcription tools. Preprocessing automatically deletes silences (ie. very low intensity signal over 300 ms duration) and cuts each raw audio file into smaller-sized files. This results in keeping around 25% of the initial input duration. In addition to each new audio file, a corresponding transcription file is created with automatically generated candidate speech turns (based on silence splits).

4.3 Transcription

The transcription was conducted by two types of transcribers: 1) students in the aeronautical field: they are not familiar with the transcription activity itself but they are highly specialized in aeronautics; 2) language specialists from translation/transcription companies, who are not especially familiar with the aeronautical phraseology but master the language. They used the free tool *Transcriber*.

All English utterances are transcribed according to American-English spelling rules. Mispronunciations are not annotated: the intended word is transcribed instead.

Besides the transcription of the pilots/controller's

[1] http://trans.sourceforge.net/

exchanges in English, we asked the transcribers to annotate additional information, listed in Table 6.

Class of info.	Type of info.	Values
Speaker	Function	pilot / controller
	Unique identifier	integer
	Native language	ISO 639-1 code
	Geo. lang. variant	ISO 3166-1 code
	Gender	male/female/unk
Spoken particularities	Gap fillers	huh
	False starts	- (e.g. *del- delta*)
	Not intellig.words	_ (underscore)
	Breaks/pauses	/
	Noise	#
	Foreign lang.	@

Table 6: Linguistic annotations

Transcriptions also contain *call sign* annotations, with distinctive annotations for full forms and short forms. Call signs are used by pilots and controllers to identify to whom they are speaking. Full forms are used when it is the first aircraft identification or when there is a risk of confusion with another aircraft identifier, ex.: *"hello* ***Ryanair*** ***nine*** ***seven papa foxtrot*** *holding november four*". Short forms are used when no confusion is possible: ***C: "Lufthansa four three uniform*** *expect to vacate via mike two"; **P: "four three uniform"***.

Other specificies of ATC are: large use of numbers and figures; use of the ICAO alphabet (*alpha* for *A*, *bravo* for *B*...); use of procedural words (*okay, wilco, roger*) and acronyms (*QNH, CAVOK*...) that are either spelled out or read; heavy use of geographical references such as waypoints or cities. Transcription homogeneity was ensured by constraining transcription guidelines and by using reference lexicons.

4.4 Quality assurance process

Formal aspects of the transcriptions, like their encoding, the syntax of annotations, the presence of unauthorized characters, etc. are checked automatically. Transcriptions shall be 100% compliant. Then, a manual check is performed by a senior linguist with expertise of ATC phraseology on randomly selected samples (50% of transcriptions). Some criteria are allowing a margin of tolerance (e.g. 1 spelling error each 10 min. of audio); other are not (e.g. 0 error on speaker features).

4.5 Post processing

Transcribed files are split into two subsets: 88% (~52 hrs for now) to be used as a train/validation dataset to develop speech recognition engines; and 12% (~7 hrs for now) to be used as an undisclosed gold-standard to benchmark speech engines at the very end of the project.

5. Application to French-accented ATC speech recognition

The goal of this section is to compare what performance may be expected with state-of-the art speech recognition techniques on French-accented ATC with: 1) AIRBUS-ATC data only; 2) NIST corpus only (real-life ATC, mainly US-English accented); 3) AIRBUS-ATC data combined with NIST corpus.

The experiments were conducted on the train/validation part of the AIRBUS-ATC corpus. 80% of the train/validation corpus (~42 hrs) was used to train a state-of-the-art engine (alone or in combination with NIST). The remaining 20% (~10 hrs) were used to evaluate the models (see WER results in Table 7).

The speech recognition engine uses state-of-art techniques. The acoustic model is a Time Delay Deep Neural Network (TDNN) (Peddinti, 2015) containing 6 layers with 4 hidden layers and a total of 6.1M parameters. The language model (LM) is a 4-gram LM built from the transcriptions. We used CMU-Sphinx[2] dictionary and added pronunciation of all the words missing from the dictionary but present in the transcriptions.

Results in table 7 tend to support the claim that accent-specific corpora are key to obtaining good performances on accented speech.

We hypothesize that poor performance of the NIST corpus is largely due to the language model: NIST corpus is older than AIRBUS-ATC corpus (NIST has been recorded in 2004); US ATC terminology differs from the EU terminology that has to follow ICAO (International Civil Aviation Organization) rules; and AIRBUS-ATC also has a lot of lexical items in common with the test corpus (geographical references like waypoints, cities, specific airlines call signs, etc.). Regarding the 0.3% WER decrease of the AIRBUS-ATC+NIST combination, we hypothesize that this is due to the acoustic model performing better trained on both data sets than each of the sets separately.

Training data	Volume	Word Error Rate
NIST	70 hrs	94.7%
AIRBUS-ATC	42 hrs	12.7%
AIRBUS-ATC + NIST	112 hrs	12.4%

Table 7: Performance of the ASR systems

6. Lessons learned

As with all spoken language corpora, constituting an ATC corpus is a complex and tricky task. We faced technical challenges like the one during the recording phase (§4.1) but also methodological issues. The transcription task is

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

undoubtedly time-consuming and a costly investment.

6.1 Transcription guidelines and training

We initially proposed complex high precision transcription guidelines, with a lot of expected annotation details. Indeed, we asked transcribers to make distinctions between instantaneous and lasting events (silence or noise), to be able to distinguish what belongs to ICAO phraseology or not, to highlight bad pronunciation, human noises like breathing, etc. All in all, there were more than 45 rules and this was too complex both to transcribe and to check. Consequently, we decided to optimize this part of our process by a strong simplification of the transcription guidelines while respecting the objectives and the expected quality of the transcriptions.

Moreover, we underestimated the importance of specifically training transcribers to this task. As we first worked in collaboration with aeronautical specialists, we thought that simply giving the transcription guidelines and a short briefing would be sufficient to get quality transcriptions. But, the first results were not only transcribed with a lot of spelling errors but also delivered later than expected: we misevaluated the time needed for transcription. In a second phase, we then contacted language specialists. We strongly insisted on the training, with face-to-face workshops on what is expected in terms of transcription and annotations and with continuous support. This greatly improved the quality of transcriptions.

6.2 Issues with spoken language specificities

For most transcribers, a major difficulty lies in discriminating repeated, broken words and lasting syllables. In fact, a repeated word with a specific lasting accentuation on the last syllable was often transcribed as a broken word. In the examples below we underline the lasting syllable for a better understandability; the hyphen ("–") is the convention used to indicate a broken word:

1. *continue- continue*
2. *wind is one two zero degrees two- two zero knots*

In both examples, the first word of the repeated sequence is a complete word, and not part of a word although the hyphen should correspond to a break before the end of a word. Expected transcriptions are:

1. *continue continue*
2. *wind is one two zero degrees two two zero knots*

The identification of a speaker's characteristics like native language (and geographical variant) are also really difficult. It is indeed tricky for a French native to differentiate between a speaker coming from China or from Japan. Moreover, a speaker generally talks many times, in more than one non-adjacent speech turns: the transcriber needs a high level of attention and audio memory to be able to recognize and set a unique identifier to the same speaker.

Finally, code-switching between French and English words can be misleading. In the following examples, the controller starts his sentence in English and ends it with French .

> *Quality zero five four sierra huh @* → where @ equals to *"alerte huh relief vérifiez votre altitude*" (terrain obstacle check your altitude)

In the example below, a pilot utters a sentence in English except for the numbers which are enunciated in French:

> *cleared for takeoff huh three two right Quality @ hotel* → where @ equals to "*cent vingt quatre*" (hundred twenty four)

6.3 Transcription work time

	TWR-APPR	ATIS
Under trained **ATC Specialists** French speakers	20 to 40 min	20 min
Under trained Not ATC Specialists **English speakers**	60 min	N/A.
Trained Not ATC Specialists **French/English speakers**	6 to 20 min	N/A.

Table 8: Transcription duration for 1 min of audio

The figures from Table 8 show that being an ATC specialist or an English native is not particularly an advantage for the transcription of ATC communications in English. The best results are given by trained language specialists with a mixed team of native English and French people.

7. Conclusion and perspectives

This paper synthesizes the work that led to the collection of unique real-life, French-accent, speech corpus of Air Traffic Control communications aimed at building an ATC-specific speech recognition engine. Preliminary results obtained with the corpus using a state-of-the-art engine are encouraging. We also shared techniques used to collect the data, ensure quality transcription and lessons learnt from this experience. Our major perspective lies in the improvement of the linguistic resources: increasing size and accent variety of the audio data as well as developing ATC-specific pronunciation lexicons. A detailed evaluation will be performed to investigate the influence of other accents, control type (APPR-TWR vs. ATIS), speaker type (controller vs. pilot), etc. on error rate.

8. Acknowledgements

This work was funded by AIRBUS, HONEYWELL and the Clean Sky 2 research program.

10. Bibliographical References

Cauldwell, R. (2007). SpeechInAction: Fluency for Air Traffic Control. PTLC. London. August 2017. Online: http://www.phon.ucl.ac.uk/ptlc/proceedings/ptlcpaper_19e.pdf

Graglia, L., Favennec, B., and Arnoux, A. (2005). Vocalise: Assessing the impact of data link technology on the R/T channel. The 24th Digital Avionics Systems Conference, Vol. 1.

Hann, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and A. Y. Ng. (2014). Deep Speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

ICAO (2007). Manual of Radiotelephony. Doc 9432-AN/925, 4th edition.

Junqua, J-C., Fincke, S., and Field, K. (1999). The Lombard effect: A reflex to better communicate with others in noise. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 4.

Peddinti, V., Povey, D., Khudanpur, S. (2015). A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts. In proceedings of INTERSPEECH 2015. Dresden, Germany, pp. 3214-3218.

Lopez, S., Condamines, A., Josselin-Leray, A., O'Donoghue, M., and Salmon, R. (2013). Linguistic analysis of English phraseology and plain language in air-ground communication. Journal of Air Transport Studies 4.1: 44-60.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., and Picheny, M. (2017). English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:1703.02136, pp. 132-136.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D. and Zweig, G. (2016). Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256.

11. Language Resource References

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. LREC. Vol. 4, pp. 69-71.

Godfrey, J. (1994). *Air Traffic Control Complete* LDC94S14A. Web Download. Philadelphia: Linguistic Data Consortium.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 1, pp. 517-520.

Hofbauer, K., Petrik, S., and Hering, H. (2008). The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech. LREC.

Pigeon, S., Shen, W., and van Leeuwen, D. (2007). Design and characterization of the non-native military air traffic communications database (nnMATC). INTERSPEECH.

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