INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Towards an accent-robust approach for ATC communications transcription

Nataly Jahchan[1], Florentin Barbier[2], Ariyanidevi Dharma Gita[1], Khaled Khelif[2], Estelle Delpech[2]

[1]Apsys-Airbus, France
[2]Airbus, France

{nataly.jahchan, florentin.barbier, ariyanidevi.dharma-gita}@airbus.com,
{khaled.khelif, estelle.e.delpech}@airbus.com

Abstract

Air Traffic Control (ATC) communications are a typical example where Automatic Speech Recognition could face various challenges: audio data are quite noisy due to the characteristics of capturing mechanisms. All speakers involved use a specific English-based phraseology and a significant number of pilots and controllers are non-native English speakers. The aim of this work is to enhance pilot-ATC communications by adding a Speech to Text (STT) capability that will transcribe ATC speech into text on the cockpit interfaces to help the pilot understand ATC speech in a more optimal manner (be able to verify what he/she heard on the radio by looking at the text transcription, be able to decipher non-native English accents from controllers, not lose time asking the ATC to repeat the message several times). In this paper, we first describe an accent analysis study which was carried out both on a theoretical level but also with the help of feedback from several hundred airline pilots. Then, we present the dataset that was set up for this work. Finally, we describe the experiments we have implemented and the impact of the speaker accent on the performance of a speech to text engine.

Index Terms: ATC, accent robustness, speech recognition, human-computer interaction

1. Introduction

Automatic speech recognition domain in general has known several innovations leading to performance improvements. Many deep learning methods have proven their efficiency on an ASR task and today Time Delayed Neural Network (TDNN) [1], Long-Short Term Memory (LSTM) [2], [3] and Transformers [4], [5] are examples among many possible approaches to perform good results on this task. Although ATC speech constitutes a specific domain of communication with its own vocabulary and environment of recording, the results of the ATC challenge have shown that ASR technologies could be applied in this domain. While significant progress was recently observed on clear speech data, ATC transcription is still a challenge that has to be dealt with separately from natural speech transcription. Moreover, the accent consideration is still a major challenge for most of the speech recognition systems as it could be alleviated in [6], [7] or [8], and the ATC domain becomes a key element to achieving a fully operational ASR system used in aircraft cockpits. Indeed, as all actors involved on the ground and in the air come from different backgrounds and consequently have different accents when speaking English (influenced by their own native language), it is important for the ASR engine to adapt to the speaker's accent as much as possible.

This paper presents our study on the application of recent ASR approaches in the ATC domain, and introduces our investigations on accent considerations and their impact on human and automatic comprehension. We show our results obtained with HMM/TDNN approach on the Airbus ATC Challenge data [9] containing recordings registered in French airports to which we added transcribed ATC coming from Hong Kong and Taipei airports. We show the impact of the accent on the quality of the transcription according to the available data for each accent considered during the model training phase.

2. Accent analysis study

The aim of this project is to enhance pilot-ATC communications by adding a Speech to Text engine that will transcribe ATC speech into text on the cockpit interfaces to help the pilot understand ATC speech in a more optimal manner (be able to verify what he/she heard on the radio by looking at the text transcription, be able to decipher problematic English accents from controllers, not lose time asking the ATC to repeat the message several times).
In order to train the speech to text algorithm, we need to expose it to many different types of accented English. Although controllers all over the world speak English and use aviation phraseology, their speech accents are affected by their native language, and sometimes it is considerably difficult for pilots to understand controllers, which makes their task harder.

Therefore, our job as linguists and human factors specialists is to determine which types of ATC corpora, from which country/region/airport would be necessary to train the algorithm. We need to determine which accents are perceived to be the most difficult by pilots in order to provide them with a useful tool that will be able to decipher difficult accents for them.

Following an internal linguistics study into the theoretically most difficult accented English (this took into account speech rate, consonants functional load, word stress, intonation and rhythm), a socio-linguistic survey was launched. It was destined to pilots all around the world. More than a thousand pilots responded and we were able to provide the designers with a list of most problematic accents which included statistical analyses and recommendations of corpora to be taken from specific countries, regions, and airports to better train the algorithm, and eventually create a useful cockpit pilot aide.

In this paper we will focus on the hypotheses and results of the survey, and eventually compare them to the list of most problematic accents of the theoretical linguistic study in order to find common ground (refer to table 1).

http://dx.doi.org/10.21437/Interspeech.2021-333

2.1. Pilot Recruitment

Pilots were contacted through social media and were encouraged to disseminate the survey in their circles. Airbus test and training pilots and Airbus airline clients around the world also participated, however the survey was open to all pilots flying any type of aircraft. Moreover, we did not limit the study to pilots flying commercial aircraft but to all and any pilot who has had experience communicating with ATC (airline, military, private, etc.). The survey was completely anonymous (no names, emails, or IP addresses were collected). Some of the collected information for analysis and variable correlation purposes were age, gender, native language, other spoken languages, English level, flight hours, flight experience, certifications, specific regions of flight, etc.

2.2. Results and analysis

1014 pilots were recruited. 38 native languages were represented from 54 different countries. Among others: Arabic, Turkish, Persian, Hindi, Mandarin, Armenian, Afrikaans, Bengali, Urdu, Thai, Telugu, Tamil, Punjabi, Nepali, Marathi, Malayalam, Malay, Indonesian, Korean, Japanese, as well as a large range of European and South American languages. 23 of these 38 languages constitute the mother tongue of 4.1 billion people on earth (out of 7.2 billion). This represents a good coverage of the general population.

We asked pilots around which country or airspace have they had difficulties, as it would have been imprecise to ask them what accent exactly was problematic (pilots would not necessarily have known which accent is due to which native language). We assumed that most controllers of a specific country often are native speakers of the main language(s) of that same country.

Figure 1 shows the 12 accents (shown by country) which were deemed the most problematic by the surveyed pilots (difficult to understand, required several repetitions, etc.). Percentages represent the frequency of the times these countries' accents were determined to be most problematic by pilots. Note that results showed major hubs often in capital cities are always the most problematic (more traffic means more pressure to speak faster etc.), and regional accent specificities were not significantly more problematic.

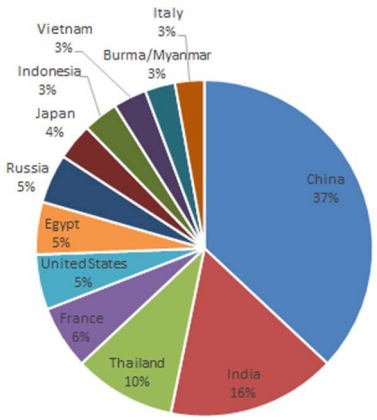

Figure 1. *Top 12 Problematic Controller Native Languages in Accented English Phraseology*

Results show that extraneous variables such as pilots' native languages as well as their level of English comprehension and flight experience did not significantly affect the perception of difficulty of a specific accent. In other words, pilots from all over the world (different native languages, different levels of English comprehension) and with a varying number of flight hours had on average the same difficulty linked to understanding specific difficult accents. Moreover, it is not because a pilot rarely encounters an accent that it is difficult to understand, rather an accent is significantly more difficult to understand despite being frequently encountered by pilots.

These results go hand in hand with the theoretical research conducted prior to this survey which investigated the perception of difficulty in accented English specificities of different language groups, speech rate, intonation and rhythm, word stress, and consonants functional load. Consequently, the difficulties arise mainly from the accent itself, influenced/constructed by the native language of the speaker (air traffic controllers all over the world) when speaking English (as a first or second language) rather than the pilot's native language influence on oral comprehension.

Table 1 offers a list of 16 countries and languages (in random order) issued from both theoretical difficulty accent study and results of this survey. Therefore, we recommend corpora taken from major city hubs in these countries for the training of the STT algorithm to cover most problematic accents around the world.

Table 1. Top *16 Recommended Problematic Accents for selecting training corpora for STT algorithm*

N°	Country and Language	Source
1	China (Mandarin Chinese and Cantonese)	Survey and theoretical study
2	France (French)	Survey and theoretical study
3	United States (English variety)	Survey and theoretical study
4	Egypt (Arabic variety)	Survey and theoretical study
5	Russia (Russian)	Survey and theoretical study
6	Japan (Japanese)	Survey and theoretical study
7	Italy (Italian)	Survey and theoretical study
8	Indonesia (Indonesian)	Survey Only
9	India (Hindi, Marathi, Telugu, Tamil, Bengali, Urdu, English variety)	Survey Only
10	Vietnam (Vietnamese)	Survey Only
11	Burma/Myanmar (Burmese and potentially other regional languages)	Survey Only
12	Thailand (Thai, or other regional varieties)	Survey Only
13	Korea (Korean)	Theoretical Study Only
14	Portugal, Brazil (Portuguese)	Theoretical Study Only
15	Spain, Latin America (Spanish)	Theoretical Study Only
16	Germany (German)	Theoretical Study Only

3. Dataset Description

Our experimental data consists of transcribed ATC speech with additional human annotated speaker metadata such as his/her accent. Since every speaker uses the same ATC English phraseology (only English speech was taken into account, the use of other languages were discarded by human transcribers), the accent annotation categorizes the speaker's accent by his/her native language (such as French, English, Spanish, etc.), and also by the variety of the language (United States, United Kingdom, Australia for the English language for example). Our corpus is based on the French-accented corpus of ATC described in [10] which is essentially composed of recordings captured at Toulouse-Blagnac Airport (LFBO) and also in other French airports (in minor proportions). This corpus has a great representation of certain European accents: a major part of it is French accented, but also includes a great amount of English, German and Spanish speakers. According to survey results in Table 1, having mainly French accents in the corpus will contribute to train and enforce an STT engine since French accents are in the top 5 most complex accents for pilots around the world.

However, in order to guarantee STT engine robustness to a wider range of problematic accents, we added to the original corpus transcriptions of ATC audio recorded in two airports from Chinese-speaking territories: the Taiwan Taoyuan International Airport (RCTP) and the Hong Kong International Airport (VHHH). These recordings have been collected on the LiveATC [1] website and the annotation follows the same guidelines as the French airport corpus. We chose these two airports to make sure we have a sufficient amount of Chinese-accented speech to train an STT engine to learn this accent (as Chinese accents were also deemed, in the survey, to be some of the most complex for pilot comprehension). In addition to the original corpus containing 100 hours of French airport ATC transcriptions, we added an extra 50 hours of Chinese-accented ATC transcriptions from Hong Kong and Taipei. This mainly added Chinese, English, and a small amount of French accents to the corpus.

3.1. Accent distribution in the dataset

The accent distribution is given in Table 2 for both the training and test set.

Table 2. *ATC corpus accent distribution*

Accent	% of train set	% of the test set
French	37.76	27.23
German	3.23	2.67
Spanish	2.84	1.47
English	20.34	22.86
Chinese	13.43	20.66
Unknown	18.5	21.57
Others	3.74	3.54

We built the test set on the Airbus ATC Challenge evaluation set [9] which contains a 5-hour dataset extracted from the same airports as the training set. It contains approximately the same distribution of accents. To also evaluate the performance of our engine on the accents we could encounter at airports in Chinese-speaking countries, we selected 10% of the 50 hours of Chinese ATC corpus while respecting the ratio of accents present in the original Hong Kong-Taipei dataset as a whole.

To simplify the accent categorization of the speakers of our dataset, we regrouped the accent by mother tongue instead of keeping all the varieties of French, German, etc. It gives us a more representative distribution of the accents in the corpus. As stated before, the most present accent is the French accent, followed by the English accent and the Chinese accent, which thanks to our additional data is now significantly present in the corpus. The "Unknown" category regroups the audio where, because of the noise or the specific voice of the speaker, the accent could not be determined by transcribers. The "Others" category regroups all the identified accents having less than 1% of presence in the training corpus.

3.2. English accent in the corpus

We decided to take into consideration the varieties in the English accents as it seems plausible that an engine will not understand Australian, British and American accents in the same manner. Hence, Table 3 shows the distribution of the different varieties of English.

Table 3. *English regional distribution in the English accented data*

Accent	% in the train set English accent	% in test set English accent
En_Australian	20.13	29.01
En_UnitedKingdom	50.51	41.81
En_USA	16.22	22.38
En_Unknown	8.53	5.09
En_Others	4.62	1.71

The majority of English speakers in the dataset is British. This allows us to verify if the STT engine detects and differentiates between an American English accent and a British English accent, as the engine it was able to do between French and English ATC speech. The corpora from Chinese-speaking airports included many ATC English accents: Apart from British accents, the corpora included American and Australian accents as well.

4. Experiments

The experiments presented in this paper aim to show the efficiency of a state of the art ASR approach for ATC speech transcription, and to observe and measure its robustness to the different accents we can encounter in the ATC domain.

4.1. Model description

The comparison of the score throughout the available accents of our dataset gives an additional overview of the influence of the training on the testing of the accented speech.

We chose to experiment on the STT engine by building with the HMM/TDNN hybrid approach as described in [11] and the same loss function as described in [12] . The approach has been built with the Kaldi toolkit [13]. For the language model we used a 5 order n-gram model built with SRILM [14] computed on all the transcriptions of our dataset. Finally, we used the

[1] https://www.liveatc.net/

Sequitur G2P tool [15] to synthetize the ATC vocabulary pronunciation.

The TDNN itself is 13 layers TDNN with a dimension of 1024 and a bottleneck layer, as used in [16]. More precisely, the three first layers have a time-stride of 1, the fourth a time stride of 0, and the following eight have a time-stride of 3.

As input, we used 40 dimensioned Mel-Frequency Cepstral Coefficients (MFCC), and we used the additional i-vector feature [17] coming from an extractor train on our training data.

4.2. Experiments results

To measure the performance of the engine throughout the available accents, we used the Word Error Rate (WER) computed by the script given for the Airbus ATC Challenge. We scored the performance on the whole test set and also on the Airbus ATC Challenge test set. The Chinese speaking Airports test set was separated. These scores are listed in Table 4. The results show an affinity of the STT engine for the ATC Challenge test set. Indeed, as most of the training data are issued from the same airports (2/3) as this test set, itis logical that the score is far better for this corpus. However, the engine succeeded to learn additional accents brought by the Chinese airports corpora.

Table 4. *WER by test set - FR-ZH airports STT engine*

Test set	WER (%)
ATC Challenge test set	9.24
Hong Kong/Taipei test set	20.33
Both datasets	14.43

Indeed, as the addition of the Chinese corpora have mainly added Chinese and English accented data, we have to take into account the proportion of these accents in the training dataset to evaluate the performance of the engine on them. Moreover, the English-accented data brought by the Hong Kong/Taipei dataset are not only British accents unlike the ones coming from our first corpus. As we kept the same proportion of accents in the Hong Kong/Taipei test set, the tested Australian and American English accented data are not as present as the British accents, which have to be taken into account when considering the performance on these accents.

Table 5. *WER computed by accent - FR-ZH airports STT engine*

Accent	WER (%)
French	9.33
German	10.58
Spanish	12.62
English	16.95
Chinese	15.55

When looking at the per-accent WER score of the engine in Table 5, we can see that scores are different for each accent. The French accent has the best score. Considering the amount of French accented data in the training data, it is normal that the best transcribed accent by the engine is the French speech. When evaluating all the scores, we can observe that the engine does not understand each accent with the same accuracy. Apart from German and Spanish (not enough data tested to take them into account), we can see that Chinese languages have a better WER than English, while it is less present in the training set. Our hypothesis is that the engine has trouble learning all the varieties of English accents in the training data. Table 6 shows the performance of the engine for the different varieties of English accents.

Table 6. *WER computed by English accent varieties - FR-ZH airports STT engine*

English accent	WER (%)
En_Australian	15.55
En_United Kingdom	14.77
En_USA	18.24

The STT engine does not have the same transcription capability for the three English accents shown in this table. The best score comes from the most present accent in the training set (British) but is barely superior to the score of the Australian accent. We can also observe that, even though the American accent is equally present in the corpus as the Australian one, the engine does not reach the same performance for the two accents. We can conclude that, based on the classification in Table 1, if the American accents are some of the most problematic accents for pilots, they are certainly equally problematic for an STT engine, and require more training data.

Finally, we removed the Chinese-speaking airports data from the training dataset and we built an STT engine only with the French airports data. This removal removed all the Chinese-accented data of the corpus, hence it cannot be evaluated with the French airports dataset alone. The Table 7 below describes the result of this experiment and confirms the impact of a difficult accent such as the Chinese one.

Table 7. *WER computed by accent - FR airports STT engine*

Accent	WER (%)
French	7.26
German	8.32
Spanish	10.04
English	8.51

5. Conclusion

The experiments were carried out using Kaldi, the well-known open-source speech recognition toolkit, and relying on the AIRBUS-ATC dataset, a real-life corpus containing about 100 hours of VHF-quality audio, along with their manual transcription. Our automatic transcription of the evaluation dataset achieved a very promising Word Error Rate (WER) despite all the challenges of this ASR task. The study on accents shows pilots have more trouble to understand certain accents, and we show with our experiment that an STT engine is able to learn a wide range of accents to mitigate this comprehension problem. Although having a wide range of accents in ATC speech certainly brings an added layer of difficulty for training an STT engine and considerably increases WER scores, we have shown that our engine still shows competitive transcription performances on the whole, and would potentially provide a helpful tool for pilots experiencing difficulties with complex accents.

6. References

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, no. 3, pp. 328-339, 1989.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural computation,* vol. 9, pp. 1735-80, 1997.

[3] A. Graves, S. Fernandez and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition," *Artificial Neural Networks: Formal Models and Their Applications -- ICANN 2005,* pp. 799--804, 2005.

[4] V. Ashish, N. Shazeer, P. Niki, U. Jakob, J. Llion, N. G. Aidan, L. Kaiser and P. Illia, "Attention is All you Need," *ArXiv,* vol. abs/1706.03762, 2017.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 8, pp. 1240-1253, 2017.

[6] J. Abhinav, U. Minali and J. Preethi, "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning," *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India,* pp. 2454-2458, 2018.

[7] D. Vergyri, L. Lori and J.-L. Gauvain, "Automatic speech recognition of multiple accented English data," *Eleventh Annual Conference of the International Speech Communication Association,* 2010.

[8] M. T. Turan, E. Vincent and D. Jouvet, "Achieving multi-accent ASR via unsupervised acoustic model adaptation," *INTERSPEECH 2020,* 2020.

[9] T. Pellegrini, J. Farinas, E. Delpech and F. Lancelot, "The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection," *Interspeech 2019,* pp. 2993--2997, 2019.

[10] E. Delpech, M. Laignelet, C. Pimm, C. Raynal, M. Trzos , A. Arnold and D. Pronto, "A Real-life, French-accented Corpus of Air Traffic Control Communications," *Language Resources and Evaluation Conference (LREC),* 2018.

[11] P. Vijayaditya, C. Guogu, M. Vimal, T. Ko, P. Danie and K. Sanjee, "JHU ASpIRE system: Robust LVCSR with TDNNS, iVector adaptation and RNN-LMS," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* pp. 539-546, 2015.

[12] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," *Interspeech 2016,* pp. 2751-2755, 2016.

[13] D. Povey, G. Arnab, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz and G. Stemmer, "The kaldi speech recognition toolkit," *IEEE 2011 workshop,* 2011.

[14] A. Stolcke, "SRILM - an extensible language modeling toolkit," *INTERSPEECH,* 2002.

[15] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Elsevier Science Publishers B. V.,* pp. 434-451, 2008.

[16] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," *Interspeech 2018,* pp. 3743-3747, 2018.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 788-798, 2011.

[18] C. Dugast and L. Devillers, "Incorporating acoustic-phonetic knowledge in hybrid TDNN/HMM frameworks," *Acoustics, Speech, and Signal Processing, IEEE International Conference ,* pp. 421-424, 1992.

[19] L. D. C. Dugast and X. Aubert, "Combining TDNN and HMM in a hybrid system for improved continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 1, pp. 217-223, 1994.