INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

ATCSpeech: A Multilingual Pilot-Controller Speech Corpus from Real Air Traffic Control Environment

Bo Yang[#1], Xianlong Tan[^2], Zhengmao Chen[#3], Bing Wang[^4], Min Ruan[^5], Dan Li[^6], Zhongping Yang[%7], Xiping Wu[#8], Yi Lin[#9]*

[#]National Key Laboratory of Air Traffic Control Automation System Technology, College of Computer Science, Sichuan University, China
[^]Southwest Air Traffic Management Bureau, Civil Aviation Administration of China, China
[%]Wisesoft Co. Ltd., China

{[1]boyang, [3]chenzhengmao, [9]yilin}@scu.edu.cn, [2]caactxl@sina.com,
{[4]kele4808, [5]minerruan}@163.com, {[6]e_dandan, [7]yzpping, [8]wuxipingstar}@126.com

Abstract

Automatic Speech Recognition (ASR) technique has been greatly developed in recent years, which expedites many applications in other fields. For the ASR research, speech corpus is always an essential foundation, especially for the vertical industry, such as Air Traffic Control (ATC). There are some speech corpora for common applications, public or paid. However, for the ATC domain, it is difficult to collect raw speeches from real systems due to safety issues. More importantly, annotating the transcription is a more laborious work for the supervised learning ASR task, which hugely restricts the prospect of ASR application. In this paper, a multilingual speech corpus (ATCSpeech) from real ATC systems, including accented Mandarin Chinese and English speeches, is built and released to encourage the non-commercial ASR research in the ATC domain. The corpus is detailly introduced from the perspective of data amount, speaker gender and role, speech quality and other attributions. In addition, the performance of baseline ASR models is also reported. A community edition for our speech database can be applied and used under a special contract. To our best knowledge, this is the first work that aims at building a real and multilingual ASR corpus for the ATC related research.

Index Terms: automatic speech recognition, air traffic control, multilingual, ATCSpeech, speech corpus

1. Introduction

Automatic speech recognition (ASR) is always a useful interface for human-machine interaction, which is also a promising technique for air traffic control (ATC). The ASR technique can be applied in the following ATC related scenes:

a) Real-time ATC speeches can be translated by an ASR system to detect a wealth of situational context information (such as controlling intent), based on which the repetition procedure can be automatically confirmed to relieve the controllers' workload and improve the ATC operational safety [1].

b) A robotic pilot can be implemented by combining the ASR with text-to-speech (TTS), which greatly reduces the cost of training air traffic controllers [2].

c) The ASR technique plays an important role in analyzing the historical ATC speech, such as traffic operation evaluation, event detection [3].

Currently, almost all state-of-the-art ASR models are deep learning ones, in which the data distribution with respect to the speech frame and text label is fitted by large-scale labelled dataset directly. Consequently, the ASR performance highly depends on the dataset. Due to the air transportation safety and intellectual property issues, it is hard to collect real ATC speech. In addition, the domain specificities make it time-consuming and expensive to annotate sufficient samples for training a practical ASR system for ATC applications.

Compared to the common speech corpus, the ATC speech provides the following challenges for the ASR task, which should be properly considered in the research.

a) Volatile background noise and inferior intelligibility: a controller usually communicates with several pilots through a same radio frequency. Therefore, the noise model of the ATC speech is changing as the speaker changes. Moreover, the radio transmission is always an obstacle to collecting high-quality speeches.

b) Unstable speech rate: In general, the speech rate of ATC speech is higher than that of in daily life. However, it is also affected by real-time conditions and shows huge difference.

c) Multilingual ASR: In general, English is the universal language in ATC communication, whereas domestic pilots speak with controllers in local languages. Thus, the real-time ATC speech is multilingual.

d) Code switching: Code-switching is applied to eliminate misunderstanding by homonyms or near-homonyms in the speech, such as "nine->niner".

e) Vocabulary imbalance: In practice, some out-of-vocabulary (OOV) words are existing in ATC speech since speakers do not comply with the terminology strictly. Moreover, some special waypoints are rarely found in the corpus. This leads to a serious situation that the frequency of different words in the corpus is extremely unbalanced, i.e., the sample sparsity.

In this paper, we strive to create a dedicated corpus for training practical ASR systems in the ATC domain, in which

* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2020-1020

the mentioned specificities are reflected to improve the overall quality. Several baseline ASR systems are also reported in this work. Our project can be traced back to October 2016, when a runway incursion occurred in the Shanghai Hongqiao international airport, China. The post-event analysis reported that the incident was expected to be detected if the real-time pilot-controller speech can be correctly translated as the input of a safety monitoring system. This incident demonstrated the importance of ASR research and its promising application prospect in ATC domain. As a foundation of this project, we established a team of 40 people to collect and annotate the real-time ATC speech for the ASR research.

In order to advocate the 'free data' movement and make contributions to the research community in ATC, we plan to share our corpus with non-commercial institutes and researchers. As far as we know, this is the first work that aims at building a real ASR corpus for the ATC application with accented Mandarin Chinese and English Speeches. To protect air transportation safety, a community edition of our corpus (about 39-hours Chinese speech and 19-hours English speech) is opened for publicly available at this time. The access permit for the released corpus can be freely applied and must be used in accordance with a special contract strictly. If someone wants to apply for an access permit, please contact our service department. Additional data service or application of our full corpus can also be provided by other terms and contracts. The baseline ASR models with test samples are also published. The detailed information of this corpus is publicly available at: https://github.com/sculyi/ASR-Corpus.

The rest of this paper is organized as follows. The speech corpora related to common application and ATC are reviewed in Section 2. Section 3 provides the detail features of our corpus. The ASR performance of several baseline models are reported in Section 4. A short summary is in Section 5.

2. Existing ASR Corpora

It is generally known that the ASR performance highly depends on the training corpus due to its intrinsic supervised learning essence. Researchers all over the world have been striving to build available training corpora all the time. Several ASR corpora for common and ATC applications were found in the literature. Although some of them are publicly available, it is still difficult to obtain massive training samples due to the complexity of the speech signal.

For the common ASR application, the CSLT at Tsinghua University, China released a 30-hours Chinese corpus, named THCHS-30 [4]. The raw speeches were generated when speakers were reading newspapers at a silent office. The AISHELL also published two Mandarin Chinese corpora, including V1 [5] and V2 [6], which strives to transform the Chinese ASR task into an industrial scale. The Librispeech [7] published a large-scale English corpus (about 1000-hours speeches generated by reading novels), which contains clean and noisy subsets. The corpus comprises three scales: 100, 360 and 500 hours. The TED-LIUM [8] is also a popular corpus, which records the English-language TED talks. It has been updated to the third release for improving the data amount and speaker adaptation. Other ASR corpora can be found at [9], along with an ASR implementation in Kaldi [10].

For ATC applications, Delpech *et al.* reviewed several military ATC speech corpora in [11], including HIWIRE [12], nnMTAC [13]. Other datasets, VOCALISE [14] and air-ground communication [15], are also ATC related speech corpus. The ATCOSIM [16] simulated the controlling speech without the radio transmission noise. It contains about 10.7 hours data and is publicly available for all researchers. The LDC94S14A [17] is a real ATC speech collected from three US airports, about 70 hours. However, it was built in the 1990s. Recently, an ASR challenge on accented English speech was held by Airbus, and 22 teams reported their results on given dataset [18]. Although there are several ATC related corpora, most of them are monolingual, simulated and very old data or need to be purchased at a high price. Therefore, building dedicated corpus with accented multilingual speech and ATC-related elements is very important to advance the ASR application in ATC domain, this is what we strive to do in this work.

3. Data Features

3.1. Summary

Almost all the speeches in this corpus are collected from the voice record devices of real ATC systems in China. Therefore, the raw speeches are spoken in Chinese and English, and with real radio transmission noise. In our corpus, since English speech is minor language in China, a very small part of the speeches is downloaded from [19] to supply the data size of English speech, where the raw speeches are published without any transcription. In short, our database is a multilingual industrial ASR corpus in ATC domain.

The raw data is monaural speech with 8000 Hz sample rate and 16 bits sample size. Each training sample has a file pair: wave and transcription, which correspond to the input and output of an ASR model, respectively. There are about 58 hours speech data in our released corpus, which can be freely applied to a non-commercial research under a special contract. Moreover, the raw speech waveforms will be published with their transcriptions for the training, dev and test dataset.

The transcription of the corpus is manual labelled, which is a human readable sentence (Chinese character and English word). In addition, to show the ASR specificities in ATC domain, some other attributions of the raw speeches are also published with the transcriptions, which are summarized as below:

a) Speaker gender: male (M) and female (F).
b) Speaker role: pilot (P) and controller (C).
c) Speech quality: clean (C) and noise (N).
d) Flight phase: ground (GND), aerodrome tower (TWR), approach (APP) and en-route area control center (ACC).
e) Areas: indicting which airport the data is collected from. In this work, the data areas are encoded as the index.

Table 1: *Data size of the ATCSpeech corpus.*

Language	Train		Dev		Test	
	#U	#H	#U	#H	#U	#H
Chinese	43185	37.01	1200	1.04	1200	1.03
English	16392	17.80	447	0.48	438	0.48

The #H and #U denote the total duration (hours) and the number of the utterance for certain datasets, respectively.

The dataset comprises of three subsets: training, dev and test, as shown in Table 1. The data samples of each part are

randomly selected from the available samples, in which we mainly focus on the diversity of the area, speaker gender, role and speech quality. Based on our data organization, at least one sample should be selected from each folder for the dev and test set. It should be noted that the data selection procedure for dev and test dataset is executed when receiving a new batch data from data team, not for the whole samples.

3.2. Speech

In the light of the ASR challenges in ATC domain, we report the statistics of our released ATCSpeech corpus, concerning the data size, utterance, flight phase, area, speaker role, gender, speech quality and speech rate. The detail information is summarized in the Table 2. The speech quality is a subjective evaluation based on the human hearing.

Table 2: *Statistics of ATCSpeech corpus.*

Items	Chinese	English	Total
Amount			
#Number (Hours)	39.05	18.76	57.81
#Utterance	45585	17277	62862
Flight phase (Hours)			
TWR	35.74	17.50	53.24
APP	3.31	1.26	4.57
Area (Hours)			
TWR1	11.26	4.89	16.15
TWR2	13.33	4.97	18.30
TWR3	11.17	7.64	18.81
APP1	3.31	1.26	4.57
Speaker role (Hours)			
Polit	22.10	9.19	31.29
Controller	16.95	9.57	26.52
Unknown	0.009	0.004	0.013
Speaker gender (Hours)			
Male	36.70	17.08	53.78
Female	3.36	1.69	5.05
Speech quality (Hours)			
Clean	39.00	18.75	57.75
Noisy	0.052	0.014	0.066
Others			
#flight	5864	2358	8222
MoD (s)	3.08	3.91	-
SoD (s)	1.02	1.34	-
MoR (word/s)	5.17	3.34	-
SoR(word/s)	1.11	0.72	-

MoD: mean of the duration. SoD: standard deviation of the duration. MoR: mean of the speech rate. SoR: standard deviation of the speech rate.

It can be seen from the result that the released corpus contains the raw speeches collected from different flight phase and controlling center, i.e., three aerodrome towers and one approach center. The number of female speakers is obviously less than that of the male speakers due to the special duties of the air traffic controller. Since the corpus was collected from real system, the actual number of the speakers cannot be counted in detail. In the future, we plan to share more data with the speaker identification. The data amount of the pilot's speech is slightly more than that of the controller. Due to the limitation of human hearing, only a little noisy data is labelled, which is mainly caused by different speaking environments. In the released corpus, there are 5864 and 2358 flights in the Chinese and English speeches, respectively. The varieties of flight call-code indicates that our released corpus covers sufficient ATC-related information.

Analyzing the duration and rate of the speeches, the special challenges of ASR in ATC domain can also be confirmed, i.e., basically high speech rate with huge difference. As a comparison, the MoR and SoR of the open Chinese corpus THCHS30 are 3.48 and 0.47 respectively, while the measurements for the open English corpus Librispeech are 2.73 and 0.47, respectively.

3.3. Vocabulary

The speech corpus is annotated by Chinese character and English word, in which 'UNK' is for the noisy parts. In this version, there are 698 Chinese characters and 584 English words in our vocabulary. Several special English words are also in our corpus, such as the waypoint 'PIKAS', 'P127'. Moreover, there are about 60 English words in Chinese speeches, such as 'alpha', 'zulu'. Similarly, some Chinese greeting words are also in the English speeches, such as 'nihao', 'xiexie'. In general, the ASR is a sequential classification task, in which the probability of each frame belongs to a certain label class is predicted. Therefore, the class balance in the vocabulary is important to train a robust ASR model. We report the occurrence frequency of the lexicon in the Fig. 1 and Fig.2 for the Chinese and English speech, respectively. It can be seen that almost half of the words appear less than 10 times, while some of them appear up to tens of thousands of times, i.e., the label classes are extremely unbalanced in this corpus. The unbalanced vocabulary is also a key issue for the ASR task in ATC. Note that, in the test dataset, a total of two Chinese characters and 10 English words are out of the vocabulary of the train dataset.

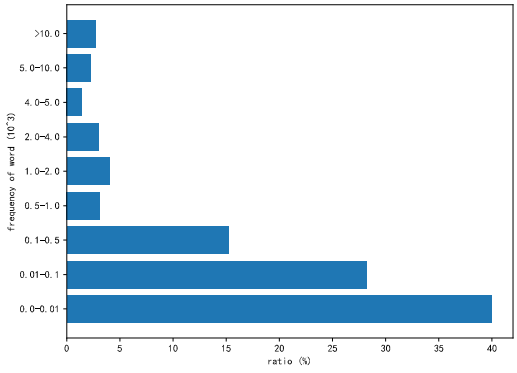

Figure 1: *Word frequency of Chinese speech.*

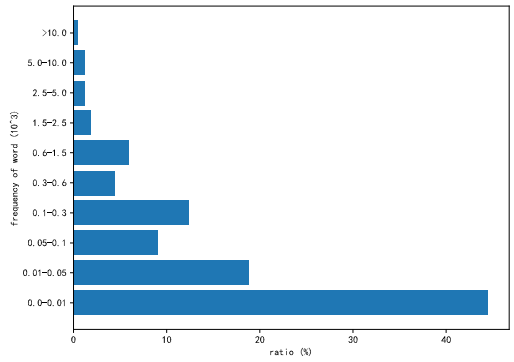

Figure 2: *Word frequency of English speech.*

4. Baseline ASR Systems

In this section, the experimental results of our baseline ASR systems are reported, which are built with the released ATCSpeech corpus. Based on the baselines, researchers can improve the model performance to address the ASR specificities in ATC domain, which is what we're trying to do to make our data publicly available.

4.1. Experiment configurations

Due to the widespread applications of the deep learning model, all the baseline approaches in this work are deep learning ones. Deep learning models are able to learn the data distributions by huge training samples and have showed state-of-the-art performance on common ASR task. All the three approaches are trained with the Connectionist Temporal Classification loss. The three baseline models are implemented by referring Deep speech 2 (DS2) [20], Japser [21], and Wav2letter++ [22], respectively. The training, dev and test dataset are same as that of in Table 1. The All baseline models are trained separately for the Chinese and English speech. The mentioned baseline models are optimized by the same data division to ensure the experimental fairness. The output vocabulary of the baseline models for Chinese speech is Chinese character and English letter, while it is English letter for English speech. A N-best (10) decoding strategy is applied to correct the spelling errors based on a N-gram character language model (LM), in which the beam width is set to 20. The LM is trained by the transcriptions in the training dataset, where a same LM is applied to all the baselines. The order of the LM are 9 and 18 for Chinese and English, respectively.

In this work, deep learning models are constructed based on the open framework Keras with TensorFlow backend. The training server is configured as follows: 2*Intel Core i7-6800K, 2*NVIDIA GeForce GTX 1080Ti and 64 GB memory with operation system Ubuntu 16.04. The training parameters for each baseline are listed in Table 3.

Table 3: *Parameters of the baseline training.*

Baseline	language	batch size	#parameter (M)	#out labels
DS2	Chinese	64	26	679
	English			28
Jasper 10*3	Chinese	32	199	679
	English			28
Wav2letter++	Chinese	64	107	679
	English			28

4.2. Baseline performance

Under certain experimental configurations, the baseline models are trained and evaluated by a same dataset. The final performance is measured by the character error rate (CER %) based on the Chinese character and English letter.

Table 4: *Baseline performance on Chinese speech.*

Baselines	**Training**		**Test**	
	Loss	Epoch	AM	AM+LM
DS2	0.53	33	8.1	6.3
Jasper 10*3	2.45	101	11.3	9.6
Wav2letter++	2.37	136	14.3	12.5

Table 5: *Baseline performance on English speech.*

Baselines	**Training**		**Test**	
	Loss	Epoch	AM	AM+LM
DS2	0.54	107	10.4	9.2
Jasper 10*3	0.91	200	9.3	8.1
Wav2letter++	1.06	307	11.3	10.1

Both the greedy decoding (AM) and beam search decoding (AM+LM) for each method are evaluated, in which experimental results for Chinese and English speech are reported in Table 4 and 5, respectively.

As seen from the experimental results, the released corpus can be applied to train all popular ASR models for both the Chinese and English speech. Overall, higher performance is achieved on the Chinese corpus Since its data size is larger than the English speech. The DS2 and Jasper based model obtain better performance for transcribing the Chinese and English speech, respectively. In addition, different baselines need different training epochs to obtain a converged accuracy depending on their model architectures. Basically, because of the recurrent neural network architecture, the DS2 based baseline needs more time to train a same data iteration, but it can be converged with less iterations. On the contrary, the training time of other two baselines are less than that of the DS2 due to its convolutional neural network architecture. However, they need more training iterations to obtain the model convergence, as shown in Table 4 and 5.

5. Conclusions

In this paper, a real multilingual ATC related ASR corpus, called ATCSpeech, is released to promote the ASR research in this filed. The raw speeches spoken in accented Mandarin Chinese and English are collected from real air traffic control systems, in which the ASR specificities are covered to improve the overall quality of the corpus. The corpus can be applied and used in non-commercial researches. Additional attributions, flight phase, speaker gender and role, and speech quality, are also summarized in the transcription. The details and CER based accuracy of baseline ASR systems are also reported in this paper. Experimental results show that all the state-of-the-art ASR models are worked on this released corpus. Based on this corpus, the ASR technique is expected to be further applied to make better decisions and continuous performance improvements in the aviation safety domain.

To the best of our knowledge, this is the first multilingual corpus that can be applied to train a practical ATC related ASR system. We sincerely hope the released corpus will benefit the new researchers and promote a promising prospect of ASR applications in ATC domain. In addition, we also hope that more ASR systems can be proposed and improved based on this dataset, which further invokes more innovation and collaboration in the research community.

6. Acknowledgements

This work was jointly supported by the National Science Foundation of China (NSFC) and the Civil Aviation Administration of China (CAAC) (Grant No.: U1833115). The authors sincerely thank all the members of the sample producing team. It is their hard work that benefits the ASR research in ATC domain.

7. References

[1] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang, and B. Yang, "A Real-Time ATC Safety Monitoring Framework Using a Deep Learning Approach," *IEEE Transactions on Intelligent Transportation Systems,* pp. 1–10, 2019, doi: 10.1109/TITS.2019.2940992.

[2] J. Ferreiros et al., "A speech interface for air traffic control terminals," *Aerospace Science and Technology*, vol. 21, no. 1, pp. 7–15, Sep. 2012, doi: 10.1016/j.ast.2011.05.002.

[3] S. Chen, H. Kopald, R. Tarakan, G. Anand, and K. Meyer, "Characterizing national airspace system operations using automated voice data processing: A case study exploring approach procedure utilization," in *13th USA/Europe Air Traffic Management Research and Development Seminar 2019*, 2019.

[4] D. Wang and X. Zhang, "THCHS-30: A Free Chinese Speech Corpus," *arXiv Preprint, arXiv1512.01882*, Dec. 2015.

[5] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5, doi: 10.1109/ICSDA.2017.8384449.

[6] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale," *arXiv Preprint, arXiv1808.10583*, Aug. 2018.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[8] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, pp. 198–208.

[9] "http://www.openslr.org/resources.php." .

[10] "http://www.kaldi-asr.org/." .

[11] E. Delpech et al., "A real-life, French-accented corpus of air traffic control communications," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2018.

[12] J. Segura et al., "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," 2007. [Online]. Available: *http://catalog.elra.info/en-us/repository/browse/ELRA-S0293/*.

[13] S. Pigeon, W. Shen, A. Lawson, and D. A. Van Leeuwen, "Design and characterization of the non-native Military Air Traffic Communications database nnMATC," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2007.

[14] L. Graglia, B. Favennec, and C. Arnoux, "Vocalise: Assessing the Impact of Data Link Technology on the R/T Channel," in *24th Digital Avionics Systems Conference*, 2005, vol. 1, pp. 5.C.2-1-5.C.2-13, doi: 10.1109/DASC.2005.1563381.

[15] S. Lopez, A. Condamines, A. Josselin-Leray, M. O'Donoghue, and R. Salmon, "Linguistic Analysis of English Phraseology and Plain Language in Air-Ground Communication," *Journal of Air Transport. Studies*, vol. 4, no. 1, pp. 44–60, 2013.

[16] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008.

[17] J. Godfrey, "https://catalog.ldc.upenn.edu/LDC94S14A," *Linguistic Data Consortium*, 1994. .

[18] 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," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*, 2019, pp. 2993–2997, doi: 10.21437/Interspeech.2019-1962.

[19] "https://www.liveatc.net/".

[20] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *arXiv Preprint, arXiv1512.02595*, Dec. 2015.

[21] J. Li et al., "Jasper: An End-to-End Convolutional Neural Acoustic Model," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*, 2019, pp. 71–75, doi: 10.21437/Interspeech.2019-1819.

[22] V. Pratap et al., "Wav2Letter++: A Fast Open-source Speech Recognition System," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6460–6464, doi: 10.1109/ICASSP.2019.8683535.