Automatic Speech Recognition Benchmark for Air-Traffic Communications

Juan Zuluaga-Gomez[1,2], *Petr Motlicek*[1], *Qingran Zhan*[1,3], *Karel Vesely*[4], *Rudolf Braun*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[4]Brno University of Technology Speech@FIT and IT4I Center of Excellence, Brno, Czechia

{juan-pablo.zuluaga,petr.motlicek,qzhan,rudolf.braun}@idiap.ch, iveselyk@fit.vutbr.cz

Abstract

Advances in Automatic Speech Recognition (ASR) over the last decade opened new areas of speech-based automation such as in Air-Traffic Control (ATC) environment. Currently, voice communication and data links communications are the only way of contact between pilots and Air-Traffic Controllers (ATCo), where the former is the most widely used and the latter is a non-spoken method mandatory for oceanic messages and limited for some domestic issues. ASR systems on ATCo environments inherit increasing complexity due to accents from non-English speakers, cockpit noise, speaker-dependent biases and small in-domain ATC databases for training. Hereby, we introduce CleanSky EC-H2020 ATCO2, a project that aims to develop an ASR-based platform to collect, organize and automatically pre-process ATCo speech-data from air space. This paper conveys an exploratory benchmark of several state-of-the-art ASR models trained on more than 170 hours of ATCo speech-data. We demonstrate that the cross-accent flaws due to speakers' accent is minimized due to the amount of data, making the system feasible for ATC environments. The developed ASR system achieves an averaged word error rate (WER) of 7.75% across four databases. An additional 35% relative improvement in WER is achieved on one test set when training a TDNNF system with byte-pair encoding.

Index Terms: Speech Recognition, Air Traffic Control, Transfer Learning, Deep Neural Networks, Lattice-Free MMI

1. Introduction

The communication methods between pilots and Air-Traffic Controllers (ATCos) have remained almost unchanged for many decades, where the ATCo's main task is to transfer spoken guidance to pilots during all flight phases (i.e. approach, landing or taxi) and at the same time providing safety, reliability and efficiency. This task has shown to be extremely stressful and highly voice demanding because of the impact a small mistake can make. Several attempts towards increasing the confidence and reducing the workload of pilot-controller communication have been pursued in the past, including experiments with Automatic Speech Recognition (ASR). Initially, due to budget and scarcity of computing power, previous work targeted isolated word recognition, or 'voice activity detection' but currently most of the works performs ASR on whole utterances. Military applications were one of the first attempts involving engines for command-related ASR; in fact, Beek et al. [1] contrast the main ASR techniques with its relevance to military applications like speaker verification, recognition of spoken codes, system control of aircraft and so on. They remarked that pilot-ATCo communications have a very limited word set -vocabulary-, speaker-dependent issues and environmental noises that need to be addressed to produce a sufficiently-reliable system. Initially, the integration of ASR technologies in ATCo started in the late 80s' with Hamel et al. report [2]; but lately, ASR technologies has been successfully deployed on ATC training simulators. For example, Matrouf et al. [3] proposed a user-friendly and robust system to train ATCos based on hierarchical frames and history of dialogues -context-dependent-. Similarly, DLR [3], MITRE [4] and more recently UPM-AENA [5] under the INVOCA project proposed akin training systems.

One of the current limitations in developing highly-accurate ASR engines for ATCo communications is the lack of available databases; likewise, generate the transcriptions of such data is extremely costly. As a matter of fact, typically a raw ATCo-pilot voice communication recording of one hour -including silences- requires between eight to ten man-hours of transcription effort [6] (mainly as it requires highly trained participants, often active, or retired ATCos). Afterwards, usually only 10 to 15 minutes speech segments of ATCo is obtained from 1h recording (after removing silence segments). Hence, it would take approximately one man-week work to get an hour of ATCos without silences [5, 6].

Currently, several researchers [7] and the International Civil Aviation Organization (ICAO) determined that the air-traffic is expected to grow about 3 to 6 percent yearly at least until 2025. Consequently, it has been seen a huge investment of the European Union (EU) to address the ATCos workload and development of ASR engines for field pilot-ATCos communication and not only for training purposes. Two recent projects financed by the EU on the scope of ASR for ATCo communications are MALORCA[1] and ATCO2[2]. MALORCA project (together with AcListant[3]) demonstrated that ASR tools can reduce ATCos workload [8] and increase the efficiency [9]. MALORCA also addressed the lack of transcribed air traffic speech data using semi-supervised training to decrease Word Error Rates (WER) and command error rates [10, 11]. We set as baseline word error rates the results from [10, 11] for two proposed train/test sets. ATCO2 ongoing project aims at developing a unique platform to collect, organize and pre-process air-traffic speech data from air space. ATCO2 considers the real-time pilot-ATCos voice communication available either directly through publicly accessible radio frequency channels (such as LiveATC [12]), or indirectly from air-navigation service providers. One of the current challenges of ASR engines for ATCo communications is the changing ATCos accent and vocabularies across different

[1] MAchine Learning Of speech Recognition models for Controller Assistance, http://www.malorca-project.de/wp/

[2] AuTomatic COllection and processing of voice data from Air-Traffic COmmunications, https://www.atco2.org/

[3] Active Listening Assistant, www.AcListant.de

airports; hence, ATCO2 will develop a robust methodology capable of minimize their impact on the system. In this work, we present the first results -or a benchmark- based on six ATC in-domain databases which, to the authors' knowledge, is the first time that such quantity of command-related databases (spanning more than 170 hours) have been used during the training phase. Firstly, we explore transfer learning from a Deep Neural Network (DNN) system trained on an Out-Of-Domain (OOD) corpus, then we contrast the results with the state-of-the-art ASR chain recipes (from Kaldi's toolkit [13]) such as TDNNF and CNN+TDNNF. Also, we concluded that there is a huge opportunity for byte-pair encoding (BPE) algorithms (used as a new representation in lexicon instead of word-based units) due to the ATCo speech-data structure i.e. the ATCo communications follows a simple vocabulary where the most spoken words are numbers. The BPE algorithms do not restrict the 'units' (in LMs, words) length and those units are not attached only to one word.

Even though obtaining a full ATCo-pilot communication system goes far beyond of only ASR tasks, we plan to convey in the following sections a benchmark of experiments going from transfer learning (from an OOD corpus) and adaptation with partial or complete in-domain command-related databases to BPE algorithms and end-to-end TDNNF models. Section 2 defines the corpus and data preparation used for our benchmark experiments. Section 3 reviews the lexicon and language modelling. The acoustic modelling and experimental setup is presented in Section 4. Then, Section 5 reviews and discusses the main obtained results. Finally, Section 6 concludes the paper and proposes the roadmap that ASR systems for ATCo communications should be heading.

2. Data Preparation

Diverse studies conclude that almost 80% of all pilot radio messages contain at least one error and 30% of the incidents are accounted by miss-communications (and up to 50% in the terminal manoeuvring area) [14]; therefore, ASR systems stand as a viable solution. Kleinert et al. [10] mention that a new technology for Air-Traffic Management (ATM) such as ASR on pilot-ATCos communication, needs to be user-friendly, comfortable and reliable enough while keeping an affordable initial cost. Accordingly, ASR systems cannot afford to be trained and tested 'on-the-fly' in real operational environment, but we are required to build the best possible system before its deployment. With this intention, we use the state-of-the-art ASR engines that are based on DNN like Time-Delay Neural Networks (TDNN) and Convolutional Neural Network (CNN). These models are known as 'data-hungry' algorithms, because state-of-the-art ASR systems need to be trained on large amount of data to achieve and acceptable operational performance. Sadly, it can be concluded that currently in the ATM world there is a lack of such databases. One of our main contribution is to solve this problem employing partly-in-domain or 'command-related' databases, retaining similar phraseology and structure but with different speakers accents; thus, helping the algorithms to achieve lower WERs.

2.1. Command-related databases

One concern that has delayed the development of a unified ASR framework for ATM globally -or at least at country level- is the vast accent's variability between ATCos from non-English speaking countries. Often, ATCos working in the same country but at different airports may have different accents (e.g. Switzerland). There is also a large variability in dictionary used across airports, as different call-signs, commands, or parameters (e.g. waypoints) can be used. Therefore, an unadapted ASR system will provide significantly worse performance due to unseen accents, Out-Of-Vocabulary (OOV) words, different recording procedures, paramenters, etc.

Table 1: *Out-of-domain and command-related databases used for transfer-learning (pre-training) and adaptation of TDNNF and CNN+TDNNF models.*

Command-related databases			
Database	**Hours**	**Accents**	**Ref**
MALORCA	13	German and Czech	[10, 11]
LDC ATCC	72.5	American English	[15]
HIWIRE	28.3	French, Greek, Italian and Spanish	[16]
ATCOSIM	10.67	German, Swiss German & French	[17]
UWB ATCC	20.6	Czech	[18]
AIRBUS	45	French	[19]
Out-of-domain databases			
Librispeech	960	Diverse English	[20]
Commonvoice	500	English subset	[21]

In order to address this issue, Table 1 presents six databases that have -or at least posses- close similarities to ATCo's speech data, accounting to nearly 180 hours (train and test sets). In fact, the phraseology and vocabularies are shared across the databases but the speakers' accent is domain-dependent. As part of our ASR benchmark for ATC, we also measured the impact of transfer learning of DNN models trained on out-of-domain databases (i.e. Librispeech and Commonvoice presented in Table 1).

Another pilot-ATCos communication concern are the errors due to OOV words and phonetic di-similarities (e.g. "hold in position" and "holding position", or, "climb to two thousand" and "climb two two thousand"). Hence, the ICAO has created a standard phraseology to reduce these errors during the communications. Similarly, Helmke et al. [22] propose a new ontology to transcribe these ATCo-pilots communications, which will harmonize the integration into the ASR systems independently from the country of origin.

2.2. Out-of-domain databases

As part of the proposed benchmark, we measured the impact of transfer learning to address the lack of in-domain databases. The idea is to pre-train models with well-known out-of-domain databases such as Librispeech [20] (960 hours) and Commonvoice [21] (500 hours English subset) and then adapt the pre-trained models using in-domain data. The final out-of-domain train set contains nearly 1500 hours of speech data (see Table 1).

2.3. Databases split

In order to measure whether the amount of data and various English accents (including variety of non-English words) of the databases influence the training process, we merged six command-related databases in three training sets as shown in Table 2. In case of ATCOSIM, we split the database (by speak-

Table 2: *ATC in-domain training and test sets.*

Name	Hours	Description
Train data-sets		
Train1	38.7	Atcosim (train) + Malorca (Vienna+Prague) + UWB ATCC
Train2	137.7	Airbus + ATCC USA + Hiwire
Tr1+Tr2	176.4	Train1 + Train2
OOD set	∼1500	Out-of-domain set: Librispeech + Commonvoice
Test data-sets		
Atcosim	2.5	20% of Atcosim train set
Prague	2.2	From Malorca set
Vienna	1.9	From Malorca set
Airbus	1	From Airbus set

ers) in a 80/20 ratio (i.e. we used 80% of data as train/validation and the remaining 20% as test set). In case of MALORCA database, it comprises two ATC approaches (collected from two ANSPs), Vienna and Prague; the initial datasets (Table 1) were already split following Table 2. As reviewed in Section 5 the performance of our methodology and developed acoustic models is evaluated on four different test sets, where features such as ATCo accent, spoken commands, airport origin and quantity of training data are varied.

3. Lexicon and Language Modelling

3.1. Lexicon

The word-list for lexicon was assembled from the transcripts of all the ATCo audio databases (i.e. Tr1+Tr2, see Table 2) and from some other publicly available resources (i.e. lists with names of airlines, airports, ICAO alphabet, etc.). The pronunciations were synthesized with Phonetisaurus [23]. The G2P (grapheme-to-phoneme) model was trained on Librispeech lexicon, and we inherited its set of phonemes. Likewise, the 'spelled' acronyms were auto-detected, and we create their pronunciations separately.

3.2. Language Modelling

We train N-gram language models using SRI-LM [24] on the transcripts of the training set Tr1+Tr2 (see Table 2). We use a tri-gram for the initial decoding and a four-gram model for rescoring. In our results (Table 3) 'LM-3' stands for the tri-gram and 'LM-4' for the four-gram model. For the BPE model we additionally trained a six-gram, identified as 'LM-6'.

4. Acoustic Modelling and Experimental Setup

All experiments are conducted using the Kaldi speech recognition toolkit [13]. We performed training on two frequently used DNN-based acoustic models. On the one hand, we train Factorized TDNN or TDNNF [25] with ∼1500 hours of OOD speech (see Table 1) and then we adapt the resulting model with three ATC command-related data-sets (see Subsection 4.1). On the other hand, we perform flat-start CNN+TDNNF training without any kind of transfer learning or adaptation; the idea behind this is to measure quantitatively whether the amount/accent of training data helps to reduce WERs. We use the standard chain LF-MMI based Kaldi's recipe for both architectures, which includes 3-fold speed perturbation and one third frame subsampling.

4.1. Conventional LF-MMI Training

Conventional LF-MMI training of TDNNF models still relies on a HMM-GMM model to build both the alignments and lattices needed during training. The HMM-GMM models are trained with only the out-of-domain databases i.e. Librispeech + Commonvoice. We prepare 100-dimensional i-vector features, 3-fold speed perturbation, and lattices for LF-MMI training supervision. The TDNNF system trained on the out-of-domain training set (∼1500 hours) is tagged as 'TDNNF-B'. To measure the impact of the amount of training data on performance in the target domain, we train once with and once without transfer learning on the three different ATC train sets presented in Table 2. Models trained with transfer learning have 'TF' in the name (e.g. TDNN-TF-B). The systems without transfer learning simply are denoted according to their architectures (e.g. TDNNF, CNN+TDNNF or TDNNF-BPE).

4.2. Byte-Pair Encoding

As part of the benchmark experiments, we use Byte-Pair Encoding (BPE) [26] on the training transcripts to create a (subword) vocabulary to use for language modeling. BPE is a compression algorithm which transforms whole words into 'units' of sub-strings, allowing the representation of an open vocabulary where new words can be easily introduced in the lexicons and LMs. There have been several studies using BPE for ASR systems [27, 28, 29], we believe there is an especially strong case for it for ATC communications, as it relies mostly on simple commands and call-signs (our ATC vocabulary is smaller than 10k), but at the same time contains a relatively high amount of foreign proper nouns, which could be missing in a word-based model. For BPE training we limited the number of merges to 2000 (resulting in 2000 sub-words), we used the original implementation from [26]. We use a character-based sub-word lexicon which means to get a pronunciation for a word we simply split the word up into its characters, and then use these characters instead of phones. As mentioned previously, the LM is a six-gram language model. After decoding, the words that end with the separator symbol are joined with the next one, so that we end up with words as the final output and on which we can calculate the WER (comparable with word-based models).

5. Results and Discussion

The results (seen in Table 3) are split into four blocks. First, a system (TDNNF-B) is trained on an OOD set consisting of 1500 hours. This is our base model to perform transfer learning. Second, we use the TDNNF-B model to adapt to the different ATC datasets (by training on them and using TDNNF-B as initialization) i.e. Train1, Train2 and Tr1+Tr2. Third, we compare WERs for TDNNFs without transfer learning trained on each of the three proposed training sets. Finally, we present results on a CNN+TDNNF chain model and a TDNNF model trained with BPE units (see BPE section for details on the setup). We kept the same hyper-parameters across all the experiments in order to make fair comparisons between models.

The base model performs poorly on the ATC data. This is not surprising as Librispeech and Commonvoice are both read speech with mostly clear audio. The ATC data is more noisy,

Table 3: *DNN benchmarks with different training methodologies and amount of in-domain and out-of-domain training data. TDNNF-B is our proposed base model trained on Librispeech and Commonvoice. TDNNF-TF-B uses TDNNF-B for initialization (acting as transfer learned model) and then is adapted on the corresponding dataset seen in the table. TDNNF-BPE is a byte-pair encoding system based on 2k sub-word units and a 6-gram language model developed using Tr1+Tr2 train set. CNN+TDNNF is composed of six convolutional layers coupled with nine TDNNF layers at the top. These models as well as TDNNF are just trained on the displayed dataset (in the same row).*

			Word Error Rates (WER) % - (test sets)							
System	**Train Set**	**Params**	**Vienna**		**Prague**		**Airbus**		**Atcosim**	
			LM-3	LM-4	LM-3	LM-4	LM-3	LM-4	LM-3	LM-4
TDNNF-B	OOD set	23.1M	95.8	95.8	47.6	43.3	80.6	77.5	67.5	63.4
	Train1		7.6	7.1	9.1	9.0	53.6	51.4	7.5	7.3
TDNNF-TF-B	Train2	20.8M	30.2	26.2	19.3	17.8	14.9	14.6	23.9	20.5
	Tr1+Tr2		7.5	6.9	8.6	8.4	15.2	14.7	5.9	6.0
	Train1		8.1	7.5	8.9	8.7	67.8	66.7	8.5	8.1
TDNNF	Train2	20.8M	33.2	30.2	20.1	18.8	14.6	14.5	23.4	19.6
	Tr1+Tr2		**7.1**	**6.6**	8.1	7.9	**14.6**	**14.4**	5.3	5.2
CNN+TDNNF	Tr1+Tr2	14.3M	7.1	6.7	8.1	7.9	15.1	14.7	**5.0**	**5.1**
			LM-6		LM-6		LM-6		LM-6	
TDNNF-BPE	Tr1+Tr2	20.8M	7.6		**5.1**		15.1		7.2	

the speakers talk much quicker, and the accents are stronger. Despite the significant difference in domains, the pretraining still helps when the target dataset is not too large, as can be seen when comparing the first two rows of Table 3 (trained on Train1, Train2) of the TDNNF-TF-B and the TDNNF models. However, once the target domain dataset becomes large enough, we do not see the benefit of pretraining (see the last row of the TDNNF-TF-B and the TDNNF models).

The main purpose of the last block of experiments is to provide a broader cover of different DNN architectures and techniques on our proposed ASR benchmark for air traffic communications. There is no clear winner. The CNN+TDNNF system yielded a new baseline of 5% WER for Atcosim, showing a relative improvement on WERs of 16.7% and 3.9% when compared to TDNNF-TF-B and TDNNF. For the Vienna approach, our best model was TDNNF trained on Tr1+Tr2 and scored with a 4-gram LM; for Prague approach the best performing model was TDNNF with 6-gram and lexicon based on BPE. Compared to previous experiments on MALORCA [10, 11] our approach yields 29.8% and 37.9% relative WER improvement for Vienna and Prague.

We further investigated why the BPE model does significantly better on the Prague test set, and found that the difference in performance is entirely explained by reduced deletions (five times more deletions of TDNNF and CNN+TDNNF than TDNNF-BPE system). The word-based model is obviously not able to recognize out-of-vocabulary words, which is the primary reason for the deletion errors. We checked OOV rates, and found that on the Prague, Vienna, Airbus and Atcosim test sets they are 3.3%, 1.1%, 0.0% and 0.1%. This shows that the BPE system is capable of recognizing OOVs and thereby improving performance, although it does come at a cost (since the BPE models also perform significantly worse on some test sets). Further investigation is required to understand the differences in performance between word and sub-word (BPE) based systems. For instance, we noticed that the BPE model does better on foreign words (even when the word-based model includes these words in its lexicon), which we attribute to the character-based lexicon generalizing better to foreign languages which are not closely related to English.

The Atcosim baseline WER is presented in [30]. They achieved 8.5% absolute WER when performing n-best list reranking using syntactic knowledge. In our case, we obtain first 63.4% WER with TDNNF-B and an improvement to 8.1% absolute WER when training only on Train1 set. An additional 10% relative WER improvement can be obtained if employing transfer learning (i.e. TDNNF-TF-B + Train1), reaching 7.3% absolute WER. As on previous test sets, an increasing amount of training data helped the models to generalize better; consequently, we achieved an additional 28% relative WER improvement when training TDNNF on Tr1+Tr2. Finally, with the intention to explore different DNN architectures we were able to further reduce the absolute WER to 5.0% when using a CNN+TDNNF system trained on Tr1+Tr2, accounting to 3.8% relative improvement from TDNNF.

6. Conclusions

The main intention of this work is to introduce state-of-the-art DNN architectures to the area of ASR for air-traffic communications. We performed a benchmark with different DNN architectures, amount of training data and transfer learning across the presented experiments in order to reasonably compare their performance. To the author's knowledge, this is the first paper employing six air-traffic command-related databases spanning more than 176 hours of speech data that are strongly related in both, phraseology and structure to ATCos-pilots communications, therefore dealing with the burden of lack of databases that many previous studies have quoted. Specifically, we have shown that using in-domain ATC databases, even if not from the same country/airport, the system is capable to yield a 29.8% and 37.9% relative WER improvement for Vienna and Prague approaches. Also, we reported new baselines for Vienna, Prague and Atcosim test sets. Finally, one of the main outcomes of this research was the results on byte-pair encoding with Prague approach, reaching 5.0% WER. We advise that future research should be focused in this way of AM, LM and lexicon modeling.

7. Acknowledgements

The work was supported by by European Unions Horizon 2020 project No. 864702 - ATCO2, which is a part of Clean Sky Joint Undertaking. Karel Vesely was also supported by Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X.

8. References

[1] B. Beek, E. Neuberg, and D. Hodge, "An assessment of the technology of automatic speech recognition for military applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 310–322, 1977.

[2] C. J. Hamel, D. Kotick, and M. Layton, "Microcomputer system integration for air control training," Naval Training Systems Center, Orlando FL, Tech. Rep., 1989.

[3] K. Matrouf, J. Gauvain, F. Neel, and J. Mariani, "Adapting probability-transitions in dp matching processing for an oral task-oriented dialogue," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 569–572.

[4] R. Tarakan, K. Baldwin, and N. Rozen, "An automated simulation pilot capability to support advanced air traffic controller training," in *The 26th Congress of ICAS and 8th AIAA ATIO*, 2008, p. 8897.

[5] J. Ferreiros, J. Pardo, R. De Córdoba, J. Macias-Guarasa, J. Montero, F. Fernández, V. Sama, G. González *et al.*, "A speech interface for air traffic control terminals," *Aerospace Science and Technology*, vol. 21, no. 1, pp. 7–15, 2012.

[6] J. M. Cordero, M. Dorado, and J. M. de Pablo, "Automated speech recognition in atc environment," in *Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems*, 2012, pp. 46–53.

[7] H. Holone *et al.*, "Possibilities, challenges and the state of the art of automatic speech recognition in air traffic control," *International Journal of Computer and Information Engineering*, vol. 9, no. 8, pp. 1940–1949, 2015.

[8] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing controller workload with automatic speech recognition," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 2016, pp. 1–10.

[9] H. Helmke, O. Ohneiser, J. Buxbaum, and C. Kern, "Increasing atm efficiency with assistant based speech recognition," in *Proc. of the 13th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, USA*, 2017.

[10] M. Kleinert, H. Helmke, G. Siol, H. Ehr, A. Cerna, C. Kern, D. Klakow, P. Motlicek, Y. Oualil, M. Singh *et al.*, "Semi-supervised adaptation of assistant based speech recognition models for different approach areas," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018, pp. 1–10.

[11] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proc. of the 18th Annual Conference of the International Speech Communication Association*, 2017.

[12] LiveATC, "Liveatc.net - live air traffic," 2020. [Online]. Available: https://www.liveatc.net/

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[14] C.-M. Geacăr, "Reducing pilot/atc communication errors using voice recognition," in *Proceedings of ICAS*, vol. 2010, 2010.

[15] J. Godfrey, "The Air Traffic Control Corpus (ATC0) - LDC94S14A," 1994. [Online]. Available: https://catalog.ldc.upenn.edu/LDC94S14A

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

[17] K. Hofbauer, S. Petrik, and H. Hering, "The atcosim corpus of non-prompted clean air traffic control speech." in *LREC*, 2008.

[18] L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 449–464, 2019.

[19] 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," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[20] 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)*. IEEE, 2015, pp. 5206–5210.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[22] H. Helmke, M. Slotty, M. Poiger, D. F. Herrer, O. Ohneiser, N. Vink, A. Cerna, P. Hartikainen, B. Josefsson, D. Langr *et al.*, "Ontology for transcription of atc speech commands of sesar 2020 solution pj. 16-04," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018, pp. 1–10.

[23] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Nat. Lang. Eng.*, vol. 22, no. 6, pp. 907–938, 2016.

[24] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[25] 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." in *Interspeech*, 2018, pp. 3743–3747.

[26] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *ArXiv*, vol. abs/1508.07909, 2016.

[27] J. Drexler and J. Glass, "Subword regularization and beam search decoding for end-to-end automatic speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6266–6270.

[28] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *CoRR*, vol. abs/1805.03294, 2018. [Online]. Available: http://arxiv.org/abs/1805.03294

[29] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, C. E. Siong, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," in *INTERSPEECH*, 2019.

[30] H. Holone *et al.*, "N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in air traffic control," in *2016 16th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2016, pp. 1309–1314.