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

Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances

Oliver Ohneiser[1], *Saeed Sarfjoo*[2], *Hartmut Helmke*[1],
Shruthi Shetty[1], *Petr Motlicek*[2], *Matthias Kleinert*[1], *Heiko Ehr*[1], *Šarūnas Murauskas*[3]

[1]German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany
[2]Idiap Research Institute, Martigny, Switzerland
[3]State Enterprise "Oro navigacija" (ON), Air Navigation Service Provider of Lithuania, Lithuania
{firstname.lastname}@dlr.de, {firstname.lastname}@idiap.ch, murauskas.s@ans.lt

Abstract

The maturity of automatic speech recognition (ASR) systems at controller working positions is currently a highly relevant technological topic in air traffic control (ATC). However, ATC service providers are less interested in pure word error rate (WER). They want to see benefits of ASR applications for ATC. Such applications transform recognized word sequences into semantic meanings, i.e., a number of related concepts such as callsign, type, value, unit, etc., which are combined to form commands. Digitized concepts or recognized commands can enter ATC systems based on an ontology for utterance annotation agreed between European ATC stakeholders. Command recognition (CR) has already been performed in approach control. However, spoken utterances of tower controllers are longer, include more free speech, and contain other command types than in approach. An automatic CR rate of 95.8% is achievable on perfect word recognition, i.e., manually transcribed audio recordings (gold transcriptions), taken from Lithuanian controllers in a multiple remote tower environment. This paper presents CR results for various speech-to-text models with different WERs on tower utterances. Although WERs were around 9%, we achieve CR rates of 85%. CR rates only slightly decrease with higher WERs, which enables to bring ASR applications closer to operational ATC environment.

Index Terms: speech recognition, speech understanding, command recognition rate, air traffic control, tower utterances

1. Introduction

Automatic speech recognition (ASR) in air traffic control (ATC) existed decades ago [1],[2]. However, it got more powerful in the last decade due to improved computing power for model training and accelerating digitization in the ATC domain. Normally, the step that follows ASR is language understanding – in ATC, also called as spoken instruction understanding [3]. Different projects have shown possible applications [4] such as runway incursion detection [5], decision support input [6], radar label maintenance [7],[8], etc., which ultimately results in benefits such as workload reduction for air traffic controllers [9]. For language understanding, multiple words are analyzed to extract the semantic meaning (concept extraction) of utterances, which includes the extractions of ATC concepts, such as callsigns, command types, command values, units, conditions, etc. The extraction of these ATC concepts is supported by machine learning algorithms [10]. The ATC concepts can be annotated by applying the rules of an ontology, agreed by 14 European air navigation service and system providers [11]. Concept extraction has already been applied to ATC utterances from the approach domain and to manually transcribed (gold) ATC utterances from the tower domain [12]. Our approach in this paper is among the first applications to apply command recognition on partly erroneous recognized speech text from the tower domain[1]. With this approach, we investigate the effect of using unsupervised data for training a robust acoustic model for the ATC domain. The improvement of word error rate (WER) and the partly dependent enhancement of command recognition rate (CRR) are important steps to achieve higher technology readiness levels because the ATC end users are interested in low error rates on semantic level. The next section presents related work on language modeling, transcription rules, and the annotation ontology. Section 3 describes the ATC concept extraction to recognize commands as well as trials for data acquisition and analysis. The recognition experiments and results are shown in section 4. Section 5 concludes and gives an outlook on future work.

2. Related Work

2.1. Language Modeling

Several LM adaptation or interpolation techniques were proposed for mapping the language model (LM) to the specific domain, e.g., linear interpolation, Bayesian interpolation and count merging. Bayesian interpolation was introduced in [13]. [14] and [15] showed that count merging with two data sources is a specific style of maximizing a posteriori (MAP) adaptation. [16] shows the theoretical connections between the mentioned LM interpolation techniques.

2.2. Transcription Rules and Annotation Ontology

Different transcription rules for ATC utterances have been defined and used for existing audio corpora [17]-[20] such as:

- Spelled letters – not pronounced using the International Civil Aviation Organization (ICAO) alphabet such as alfa, bravo, etc. – e.g., "~k~l~m"/"KLM"/"K L M",
- Truncated/broken word parts, e.g., "luf="/"luf*"/"luf-" if "lufthansa" was not uttered fully till the end,
- Non-understandable words ("[unk]" / "[UNKNOWN]") and human noise/thinking loud ("[hes]" / "[HNOISE]"),
- Non-English words, e.g., "<FL></FL>" / "[NE][/NE]".

[1] For funding information please refer to [38],[11],[30].

http://dx.doi.org/10.21437/Interspeech.2021-935

Also, for the annotation of semantic meanings of the ATC transcriptions different ontologies or rule sets exist. An early ontology developed by NATS for the terminal environment comprised of callsign, standard type, non-standard type, value, and type unit [21]. Similarly, the ontology introduced by the *AcListant*® project [22] proposed to use four different elements: callsign, type, value, and unit of a command [23],[24]. A further approach suggested to use keywords like callsign, flightlevel, altimeter for the corresponding values [25]. Another proposition was to have ten class labels for annotation of word sequences such as callsign, fix, number, etc. [26],[27]. The *AcListant*® ontology was enhanced during the *MALORCA* project [28] in which various command types for "information", "reports", and "expects" were added next to conditional clearances [29]. This ontology has been further enhanced for en-route and tower commands during the *CWP HMI* project [11]. Furthermore, the ontology with more than 100 different command types has been agreed between major European partners from the air traffic management (ATM) domain including air navigation service providers, ATM system providers, and the coordinating partner DLR. The *HAAWAII* project [30] further enhanced the ontology for pilot utterances including their requests and reports. Also, other European ASR projects such as *HMI Interaction modes for Approach control*, *HMI Interaction Modes for Airport Tower*, and *Safety and Artificial Intelligence Speech Recognition* continuously contribute to the improvement of the ontology. The global scheme for each instruction to annotate ATC utterances is shown in Figure 1. Each ATC utterance can contain multiple instructions.

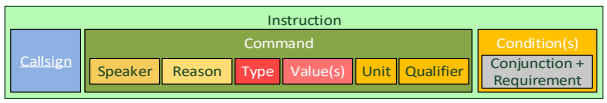

Figure 1: *Elements of an air traffic control instruction including the ATC concepts 'callsign', 'command' with sub-elements, and optional 'conditions'.*

The callsign is a mandatory element for each instruction and might be NO_CALLSIGN if not uttered. This is followed by a mandatory command and may be followed by optional conditions. The command again can have a speaker (PILOT or empty for default air traffic controller), a reason (REPORTING, REQUEST or empty), a type (REDUCE, DESCEND, VACATE, CONTACT_FREQUENCY, CLEARED VIA, etc.), one or multiple values ("200", "A B D1", "118.300", etc.), a unit (FL, ft, kt, none, etc.), and a qualifier (RIGHT, OR_LESS, etc.). The conditions have a conjunction and a requirement ("UNTIL 4 NM FINAL", "WHEN AIRBORNE", etc.). An ontology for annotations supports different purposes. It is needed as an interface to enable interoperability of different ASR applications with ATC systems. It is also necessary for evaluating automatically recognized commands against manual (gold) annotations. The name "command recognition rate" (CRR), taken from [6] has historical reasons. According to Figure 1, the term "instruction error rate" would be correct. For the calculation of the CRR, each command, for example consisting of the ATC concepts callsign, type, value, qualifier, condition, etc. is considered as one (big) word to compute the Levenshtein distance [31]. This means that a recognized command is correct only if all concepts (command parts) are correct, i.e., "DLH7HT HEADING 360 LEFT" and "DLH7HT HEADING 360 none" are not equal and would be counted as a full command recognition error. The CRR is defined as the number of controller commands correctly recognized by the ASR (and not rejected due to implausibility) divided by the total number of commands given or in other words: the percentage of given commands correctly shown on the controllers' display. An example transcription and resulting annotation is given in Table 1. A configuration file defines allowed values for taxiways, holding points, etc. to map "holding point three four to "HP_34" here.

Table 1: *Transcription and annotation example.*

Transcription	Annotation
[NE French] bonjour [/NE] hotel alfa charlie india zulu [unk] taxi to holding point three four via taxiway [hes] alfa runway in use three four and nex*	HACIZ TAXI TO HP_34 HACIZ TAXI VIA A HACIZ INFORMATION ACTIVE_RWY RW34

3. ATC command recognition and remote tower simulation trials

3.1. ATC concept extraction for command recognition

The command recognition algorithm consists of several steps, where different ATC concepts are extracted iteratively and put into relation to recognize them as single or multiple commands of an utterance (for more details see [10]). First, we try to extract a callsign from an ATC utterance by considering the callsign information from the available surveillance data (for controller utterances, only the first words are considered). Then, keywords or keyword sequences are extracted which initiate a command type. This step includes the extraction of a command type followed by value(s), unit, qualifier, etc. if applicable. Afterwards, we look for unmatched words in the complete utterance that correspond to non-extracted ATC concepts and we also look for command hints such as "feet" being used in an ALTITUDE command. We then search again for callsigns in the remaining unmatched words and then, we finally try to extract commands from unmatched numbers in the utterance. The above example transcription from Table 1 is reused for illustrating the algorithm here. The concept extraction model searches for the presence of any of the available predicted callsigns, e.g., AFR27C, DLH9LX, HACIZ (from surveillance data) in the utterance. The latter callsign matches here. Then, the keywords "taxi to" and the value keywords "holding point three four" as well as "via" and "taxiway alfa" lead to extraction of "TAXI TO HP_34" and "TAXI VIA A", respectively. The words "runway in use" and "three four" are extracted as "INFORMATION ACTIVE_RWY RW34". All other words ("bonjour", "[unk]", "[hes]", "and nex*") are not relevant for the command recognition algorithm example.

3.2. Trials for data recording and tower considerations

In March and December 2018 multiple remote tower trials with Lithuanian controllers from Oro Navigacija speaking accented English took place in DLR TowerLab in Braunschweig, Germany. These trials were conducted as human-in-the-loop simulations in the course of the project *CWP HMI-ASR* [32]. One controller was responsible for all the traffic from three international airports (named Vilnius (EYVI), Kaunas (EYKA), and Palanga (EYPA)) at the same time. In total, 41.4 hours with silence between different utterances aligned with radar data from the air traffic control simulation have been recorded. After deleting the inter-

utterance silence, 6.86 hours of pure speech in 3,919 audio files remain out of the trials, but only slightly more than 50% of the files have been manually (gold) transcribed and annotated. The simulation pilot utterances were not considered – only those of six tower controllers. The amount and division of labelled offline ASR data is shown in Table 2.

Table 2: *Description of transcribed audio data sets.*

Set name	# files	Duration (hours)	Average duration (sec)
all	1,993	3.6	6.6
adapt	1,399	2.6	6.8
test	594	1.0	6.1

The average duration of an utterance in this (Lithuanian) multiple remote tower environment is 6.6 seconds. This is significantly longer than for Vienna approach (4.4s) or Prague approach (5.1s) in real-life data from the *MALORCA* project. Furthermore, controllers instructed roughly 2.7 commands per utterance. Again, this is much more than 1.6 and 1.7 commands per utterance from Prague and Vienna approach from *CWP HMI-ASR* simulation runs, respectively. Also, the variation of words, i.e., the total number of different words used divided by the total number of used words is higher. The Lithuanian tower controllers used 560 different words (in total 32,484) compared to 196 different words (in total 31,436) for Vienna approach and 218 different words (in total 47,426) for Prague approach in *CWP HMI-ASR* simulation runs, respectively. Higher variation shows more free speech due to visual flight rules (VFR) traffic, e.g., vague and difficult to analyze commands like “fly heading north” would probably not be given to traffic following instrument flight rules (IFR). In addition, the number of different command types for tower ATC as modeled in the ontology is larger than for approach. Finally, the amount of available speech data for the tower domain is much less, because it is harder to record them as compared to the very high frequency receivers for approach ATC speech. All above-explained characteristics make it more challenging to automatically recognize tower commands.

4. Experiments and Results

4.1. Models and different error/recognition rates

All ASR experiments are conducted using the Kaldi speech recognition toolkit. The speech recognition acoustic model was trained on 195 hours of data from seven datasets in the ATC domain (model *Supervised baseline*). Description of the training datasets can be found in [33]. Hybrid deep neural network (NN)–hidden Markov model (DNN-HMM) with lattice-free maximum mutual information (LF-MMI) loss function was trained using alignment from Gaussian mixture models (GMM) HMM. State-of-the-art ASR chain recipes with convolutional NN–factorized time-delay NN (CNN-TDNNF) architecture from Kaldi toolkit was used for training. 4-gram[1] LM in ARPA format was trained using the same training set. For LM adaptation to the Lithuanian ATC domain, linear interpolation between the general LM and the LM from adaptation set with 0.8 and 0.2 weights was performed (model + *LM-mix*) due to the limited dataset. For improving the ASR accuracy and increasing the noise robustness of the trained model, we trained a *semi-supervised* model using 400 hours of unsupervised data from LiveATC dataset [34]. Incremental method was used for training the *semi-supervised* model [35]. We divided the unlabeled data to four 100 hours subsets. Starting from one subset, in each training iteration we added one unseen subset to the previous subsets. We extracted 86 out-of-vocabulary words including waypoints, airlines, and some local terms from the transcribed data. These words were added to the decoding graph for all experiments. The WER of the trained ASR models on test set is shown in Table 3. LM interpolation improved the WER on the test set by 9%. Effective mapping of LM using the dataset with similar phraseology pattern is one main reason for observing this improvement. In addition, including unsupervised data from ATC domain improved the ASR accuracy by 3%. It shows more robustness of the *semi-supervised* acoustic model w.r.t. the *supervised* model. Analysis on the recognition errors shows the majority of errors in the *supervised* baseline model are because of deviation of the main LM w.r.t. the in-domain data. *Semi-supervised* model reduced the recognition errors of the noisy segments and majority of the substitution errors are words with similarity in the pronunciation, e.g., "flight" and "sight".

Table 3: *Applied models (with 4-gram LM), word error rate (WER), command recognition rate (CRR), command recognition error rate (CER), and callsign recognition rate (CaRR) for tower utterances from Lithuanian controllers on test set in [%].*

Model	WER	CRR	CER	CaRR
Supervised baseline	20.8	59.0	14.1	79.5
+ LM-mix	11.8	78.4	8.2	93.8
+ Semi-supervised	8.8	84.3	7.7	96.3

The CRR in Table 3 is calculated on annotations. Thus, it can only loosely be compared to the sentence accuracy calculated on transcriptions – 1 minus sentence error rate (SER) – being used to evaluate ASR applications outside ATC domain. The CRR with gold transcription input, where a WER of 0% is assumed – compared to gold annotations is 95.8% with a command recognition error rate (CER) of 2.7%. From Table 3 we see that despite the high WER of almost 21%, a CRR of 59% is reached. With improved models, the WER decreases to roughly 12% and 9% which leads to CRRs of 78% and even 84%, respectively. As an example, the best and worst CRR per speaker were less than 5% different from the reported average using the *semi-supervised* model. A lower CRR does not really affect the workload of a controller. If there is no support by the ASR system in feeding recognized commands into the ATC system, the situation is comparable to today. Of course, higher CRRs reduce controller workload. However, if the CER increases, this results in additional workload for the controller to first recognize the error, then to delete the wrong result, and then to manually correct the wrong automatic input. A CER of 7.7% means that each thirteenth command needs to be corrected by the controller. With a higher WER of 20.8%, the number of errors is almost two-and-a-half times higher than 8.8%. The CER, on the other hand, also increases with increased WER but only from 7.7% to 14% (less than twice the number of errors). Using the baseline model, each seventh command would need to be corrected by the controller. The reason is that high WERs may lead to recognizing nothing at all on concept level, i.e., the recognized concept is rejected, because, e.g., a heading command of 733 degrees is extracted.

[1] 3-gram LM WERs were 0.2-0.6% higher than 4-gram LM WERs (only the latter reported in this paper) for the models.

The callsign recognition rate (CaRR) is also listed in Table 3. It is the most important ATC concept and can heavily influence the CRR, because the callsign is part of each command. From the perspective of an ATC application, the recognized callsign should be highlighted in the controller display to ease identifying the current communicating aircraft, and to speed up checking and correcting of recognized commands. The reference CaRR, i.e., automatic callsign extraction compared to the callsigns from gold annotation is 99%. The CaRR for the three model achieves roughly 80% to over 96%. Hence, the recognition rates for callsigns are much better than for commands in general. This is due to the usage of Assistant Based Speech Recognition (ABSR), first described in [6], which relies on using context information from the corresponding radar data.

4.2. Analysis of command recognition performance

Figure 2 shows the theoretical CERs if words from automatic transcriptions would be independent of each other given the three different WERs plus the perfect WER of 0%. It also presents the corresponding four achieved CRRs for the observed average number of six words per command.

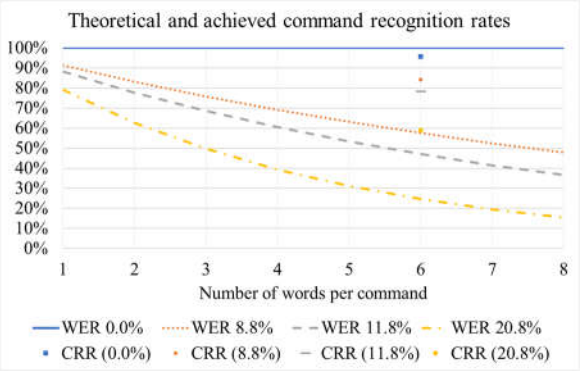

Figure 2: *Theoretical CRR for different number of words per command (sentence length) and measured CRR for average length of 6 words.*

From Figure 2 we see that with a WER of 0%, the CRR should be 100%. We, however, currently reach only 95.8% based on 68 different used ATC command types. This is on the one hand due to challenges with controller utterances being "far away" from ICAO phraseology rules [36], e.g., uttering "lufthansa two victor" if "DLH23W" is meant or instructing just the three words "two six zero", which can be a heading, speed, flight level, etc. [37]. The most recognition errors deal with command types TAXI, DIRECT_TO, and INFORMATION TRAFFIC/ACTIVE_RWY. On the other hand, the 5,367 gold annotations of the commands still contain some errors, i.e., the automatic annotation is already better than the manual annotation. However, for higher WERs, we achieve much better CRRs than theoretically achievable, if recognized words would be independent and word errors would be equally distributed. For example, a WER of 8.8% should enable a CRR of 58% for an average of 6 words per command, but we even observe above 84%. For a WER of 11.8%, we achieve a CRR of 78.4% compared to 47% based on independence assumption; and for a WER of 20.8%, we still achieve a CRR of 59% compared to 25% based on independence assumption. Hence, the WER only gives some hints to the performance of a speech recognition system in the ATC world. However, the CRR (or CER) is much more important for the end user and is more robust against higher WERs, i.e., achieve roughly 30% better recognition results than expectable due to the independence assumption.

Furthermore, it is more important to recognize longer words correctly than shorter words. If we replace each word in the speech recognition hypotheses files up to a length of 2,3,4,5,6 letters by "x", we see a steep decrease of command recognition rates when replacing words with up to three letters as shown in Table 4. If we replace the words with up to three letters, it means that we also replace the words with one and two letters. However, it is also connected to the number of replaced words, i.e., we roughly replace 0.1% (1), 5% (2), 25% (3), 51% (4), and 73% (5) of words.

Table 4: *CRRs in [%] in case of replaced words up to the listed number of letters per word (1,2,3,4,5).*

Model	1	2	3	4	5
Supervised baseline	59.0	51.6	17.1	3.0	0.4
+ LM-mix	78.4	69.6	23.6	4.9	0.6
+ Semi-supervised	84.3	75.1	27.3	5.3	0.9

This trend can be explained with the importance of certain words (given their length and number of occurrence) for the command recognition process. If words such as "a" or "A" are missing (1), there is hardly any negative effect. If words such as "to", "in", "up", "by", "or" are missing (2), there is a slight decrease in recognition. However, if meaningful words – especially numbers – such as "one", "two", "six", "via", "QNH", "KLM" are missing (3), we see a dramatic decrease. When replacing even longer words (4) such as "zero", "four", "five", "nine", "feet", "taxi", "wind", "west", "east", "left" the recognition becomes hardly usable. It is completely unusable if even longer words such as "right", "descend", "vacate", "takeoff", "knots", "degrees", "lufthansa" are replaced.

5. Conclusions and future work

This paper applies ontology-based command recognition on automatic transcriptions from ATC tower utterances of Lithuanian controllers with different WERs. Compared to the approach environment, tower utterances are longer, have more speech variety, more command types, and less available training data, i.e., recognition of words and commands is more challenging than in the approach environment. The baseline speech recognition is developed based on approach data, the first speech recognition solution uses language model adaptation, the second solution performed a semi-supervised approach leading to the best WER with around 9%. The resulting command recognition rates have proven to be robust (slight decrease) even on higher WERs. With current LM models, CRRs of 85% are possible.

In future, for alleviating the lack of transcribed speech data in the tower domain, we will focus on semi-supervised acoustic model adaptation for improving the accuracy of the ASR system on specific accents. The project *HMI Interaction Modes for Airport Tower* [38] will investigate the effect of presenting command recognition output to tower controllers in a human-in-the-loop simulation. These multiple remote tower trials will be conducted in the first quarter of 2022 in DLR TowerLab with controllers from Lithuania, Austria, and Poland. Controllers will benefit from callsign highlighting, recognized and displayed ATC concepts / commands in ontology annotation format. The presented results are already a good starting point and would enable a workload reduction compared to manually entering all given commands.

6. References

[1] D. W. Connolly, "Voice Data Entry in Air Traffic Control," Tech. Rep. N93-72621, FAA, National Aviation Facilities Experimental Center, Atlantic City, NJ, USA, 1979.

[2] C. Hamel, D. Kotick, and M. Layton, "Microcomputer System Integration for Air Control Training," Special Report SR89-01, Naval Training Systems Center, Orlando, FL, USA, 1989.

[3] Y. Lin, "Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application," Aerospace 8, No. 3: 65, 2021.

[4] J. Rataj, H. Helmke, and O. Ohneiser, "AcListant with Continuous Learning: Speech Recognition in Air Traffic Control," Air Traffic Management and Systems IV – Selected Papers of the 6th ENRI International Workshop on ATM/CNS (EIWAC2019), 6, Springer, 2021.

[5] S. Chen, H. D. Kopald, A. Elessawy, Z. Levonian, and R. M. Tarakan, "Speech inputs to surface safety logic systems," IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 2015.

[6] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, M. Kleinert, Y. Oualil, and M. Schulder, "Assistant-based speech recognition for ATM applications," 11th USA/Europe Air Traffic Management Research and Development Seminar, Lisbon, Portugal, 2015.

[7] H. Helmke, O. Ohneiser, J. Buxbaum, and C. Kern, "Increasing ATM efficiency with assistant-based speech recognition," 12th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, WA, USA, 2017.

[8] M. Kleinert, H. Helmke, S. Moos, P Hlousek, C. Windisch, O. Ohneiser, H. Ehr, and A. Labreuil, "Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance," 9th SESAR Innovation Days, Athens, Greece, 2019.

[9] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing controller workload with automatic speech recognition," IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016.

[10] H. Helmke, M. Kleinert, O. Ohneiser, H. Ehr, S. Shetty, "Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications," IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020.

[11] H. Helmke, M. Slotty, M. Poiger, D. F. Herrer, O. Ohneiser et al., "Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04," IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018. European Union's grant agreement No. 734141.

[12] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, S. Murauskas, and T. Pagirys, "Prediction and Extraction of Tower Controller Commands for Speech Recognition Applications," Journal of Air Transport Management, Elsevier, accepted 28 May 2021, expected publication June 2021.

[13] M. Weintraub, Y. Aksu, S. Dharanipragada, S. Khudanpur, H. Ney, J. Prange, A. Stolcke, F. Jelinek, and E. Shriberg, "LM95 project report: Fast training and portability," Research Note 1, Center for Language and Speech Processing, Johns Hopkins University, Tech. Rep., 1996.

[14] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'03), Vol. 1, pp. I-224–I-227, IEEE, 2003.

[15] M. Bacchiani, M. Riley, B. Roark, and R. Sproat, "MAP adaptation of stochastic grammars," Computer speech & language, 20(1), pp. 41–68, 2006.

[16] E. Pusateri, C. Van Gysel, R. Botros, S. Badaskar, M. Hannemann, Y. Oualil, and I. Oparin, "Connecting and comparing language model interpolation techniques," Interspeech, Graz, Austria, 2019.

[17] K. Hofbauer and S. Petrik, "ATCOSIM Air Traffic Control Simulation Speech Corpus," Tech. Rep., TR TUG-SPSC-2007-11, Graz, Austria, 2008.

[18] 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," Proc. LREC, Miyazaki, pp. 2866–2870, 2018.

[19] J. J. Godfrey, "Air Traffic Control Complete corpus," 1994, https://catalog.ldc.upenn.edu/LDC94S14A.

[20] S. Shetty, O. Ohneiser, F. Grezl, H. Helmke, and P. Motlicek, "Transcription and Annotation Handbook," HAAWAII deliverable D3.1, Braunschweig, Germany, 2020.

[21] D. Randall, "Direct Voice Input (DVI) Technology readiness and status introduction," Whitely, Fareham, UK, 2006.

[22] AcListant homepage: www.AcListant.de, AcListant = Active Listening Assistant, n.d.

[23] A. Schmidt, "Integrating situational context information into an online ASR system for Air Traffic Control," Master Thesis, Saarland University (UdS), Germany, 2014.

[24] Y. Oualil, M. Schulder, H. Helmke, A. Schmidt, and D. Klakow, "Real-time integration of dynamic context information for improving automatic speech recognition," Interspeech, Dresden, Germany, 2015.

[25] D. R. Johnson, V. I. Nenov, and G. Espinoza, "Automatic speech semantic recognition and verification in Air Traffic Control," IEEE/AIAA, 32rd Digital Avionics Systems Conference (DASC), East Syracuse, NY, USA, 2016.

[26] V. N. Nguyen and H. Holone, "N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in Air Traffic Control," 16th Int. Conf. on Control, Automation and Systems (ICCAS 2016), Gyeongju, Korea, pp. 1309–1314, 2016.

[27] V. N. Nguyen and H. Holone, "N-best list re-ranking using syntactic relatedness and syntactic score: An approach for improving speech recognition accuracy in Air Traffic Control," 16th Int. Conf. on Control, Automation and Systems (ICCAS 2016), Gyeongju, Korea, pp. 1315–1319, 2016.

[28] MALORCA homepage: www.malorca-project.de, Machine Learning of Recognition Models for Controller Assistance, n.d.

[29] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszák, Y. Oualil, and H. Helmke, "Semisupervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," Interspeech, Stockholm, Sweden, 2017.

[30] HAAWAII homepage: www.haawaii-project.de, Highly Automatic Air Traffic Controller Workstation with Artificial Intelligence Integration, n.d., This project has received funding from the SESAR Joint Undertaking under Grant Agreement No. 884287, under European Union's Horizon 2020 Research and Innovation programme. Idiap used funding parts for this work.

[31] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics -- Doklady 10.8, Feb. 1966.

[32] O. Ohneiser, H. Helmke, M. Kleinert, G. Siol, H. Ehr, S. Hobein, A.-V. Predescu, and J. Bauer, "Tower Controller Command Prediction for Future Speech Recognition Applications," 9th SESAR Innovation Days, Athens, Greece, 2019.

[33] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Vesely, and R. Braun, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," Interspeech, Shanghai, China, 2020.

[34] LiveATC-Homepage, https://www.liveatc.net/ , n.d.

[35] B. Khonglah, S. Madikeri, S. Dey, H. Bourlard, P. Motlicek, and J. Billa, "Incremental semi-supervised learning for multi-genre speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7419-7423, 2020.

[36] ICAO, "Doc 4444, Procedures for Air Navigation Services, Air Traffic Management," ICAO, Montréal, Canada, 2016.

[37] H. Said, M. Guillemette, J. Gillespie, C. Couchman, and R. Stilwell, "Pilots & Air Traffic Control Phraseology Study," International Air Transport Association, 2011.

[38] PJ.05-97-W2 SESAR2020 funded industrial research project under the European Union's grant agreement No. 874464, see also https://www.remote-tower.eu/wp/?page_id=888, n.d.