ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746301

CALL-SIGN RECOGNITION AND UNDERSTANDING FOR NOISY AIR-TRAFFIC TRANSCRIPTS USING SURVEILLANCE INFORMATION

Alexander Blatt[1]*, Martin Kocour*[2]*, Karel Veselý*[2]*, Igor Szöke*[2]*and Dietrich Klakow*[1]

[1]Saarland University, Saarland Informatics Campus, Germany
[2]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

ABSTRACT

Air traffic control (ATC) relies on communication via speech between pilot and air-traffic controller (ATCO). The call-sign, as unique identifier for each flight, is used to address a specific pilot by the ATCO. Extracting the call-sign from the communication is a challenge because of the noisy ATC voice channel and the additional noise introduced by the receiver. A low signal-to-noise ratio (SNR) in the speech leads to high word error rate (WER) transcripts. We propose a new call-sign recognition and understanding (CRU) system that addresses this issue. The recognizer is trained to identify call-signs in noisy ATC transcripts and convert them into the standard International Civil Aviation Organization (ICAO) format. By incorporating surveillance information, we can multiply the call-sign accuracy (CSA) up to a factor of four. The introduced data augmentation adds additional performance on high WER transcripts and allows the adaptation of the model to unseen airspaces.

Index Terms— Air Traffic Control, Call-sign Recognition, Context Incorporation, Data Augmentation

1. INTRODUCTION

The classical communication between air-traffic controllers (ATCOs) and pilots is voice-based [1]. This form of communication has the drawback, that one ACTO talks to multiple pilots over a single frequency. The rising traffic in the last years raised the number of pilots tuned in the same frequency. This increases the chance, that two pilots speak simultaneously. To avoid responses from multiple pilots, the ATCO addresses the target airplane by its call-sign. A call-sign is an unique identifier that is assigned to each airplane (e.g. `DLH83K`). New systems like controller–pilot data link communications (CPDLC), which use text based communication, reduce the load on the voice communication channels [1]. Projects like AcListant and MALORCA[1] aim to support the ATCO by speech recognition systems [2, 3]. The problem with developing such systems, is the lack of training data in the ATC domain. Although there exist some datasets [4], there is missing a database covering a multitude of locations and containing speech, transcripts and meta information like call-signs and commands. This work is part of the ATCO2 project[2], which aims among others to build up such a database.

In this work we are investigating the benefit of including context information for call-sign recognition and understanding. The context information in form of a list of surveillance call-signs is used as additional input for our models. The models recognize the call-sign in an ATC transcript and convert it to the standard ICAO format. For the training of our models, we introduce a data augmentation method, that is adjustable to the target airspace. We can show, that the models trained on the augmented data predict the target call-signs with high accuracy. We also find that the models incorporating surveillance information are superior and show a high resistance to ASR noise and surveillance data variations.

2. RELATED WORK

Various works have already investigated context incorporation in the ASR [5, 6, 7], which marks the prior step in the ATC speech processing pipeline. Two other works of the ATCO2 project [8, 9] show that the combination of HCLG and lattice boosting using Kaldi [10], reduces the ATC-ASR errors, especially for the call-signs. We build on top of these works by extracting the (erroneous) call-signs from the ASR transcripts and map them to the standardized ICAO format.

In named-entity recognition (NER) the call-sign sequence is identified in the input (Recognition), therefore it is related to our method, which additionally converts the call-sign to the target ICAO format (Understanding). NER for call-signs as single entity of interest is also part of the Airbus challenge [11]. One of the top three contestants uses a Bi-LSTM-CRF architecture [12] for the call-sign recognition, reaching an F1 score of 80.17 on the leader board. Newer pretrained transformer based models like BERT typically outperform recurrent architectures like LSTMs in natural language processing (NLP) and natural language understanding (NLU) tasks [13].

The work was supported by European Union's Horizon 2020 project No. 864702 - ATCO2.

[1]MALORCA Homepage: `https://www.malorca-project.de/`

[2]ATCO2 Homepage: `https://www.atco2.org/`

3. DATA

Table 1 contains the datasets, that are used for training and testing. The Malorca dataset [2, 3] consists of transcripts of ATCO speech from the Vienna airport together with surveillance call-signs for each transcript. The LiveATC dataset contains transcripts of ATC speech from Zurich Airport (LSZH) and Dublin Airport (EIDW) with some samples from Hartsfield–Jackson Atlanta International Airport (KATL). The speech data is collected during the ATCO2 project from LiveATC[3], which provides live ATC radio feeds. The Malorca and LiveATC transcripts are generated by three different ASR methods (baseline (b), lattice-boosting (l) and HCLG-lattice boosting (h)) [8] and by human transcription for the ground-truth data (WER 0). All transcripts are manually annotated with the correct ICAO call-sign. The generation of the augmented Airbus datatset out of the Airbus development dataset [14] is described in section 4.

Table 1. Overview of the datasets. The last column marks the WER of the different versions of the same dataset.

Datasets	Samples	WER Variants
LiveATC	500	0|28.4 (h)|28.9 (l)|33.1 (b)
Malorca	1130	0|6.42 (h)|7.27 (l)|8.47 (b)
Airbus	60000	0|7.00|30.0

A sample of the datasets consists out of the transcript (`lufthansa eight three kilo descend three thousand feet`), the corresponding target ICAO call-sign (`DLH83K`) and the surveillance call-signs (`AIF44T`, `DLH83K`, `MAN47N`, ...). The surveillance data is drawn from the OpenSky Network[4] (OSN) database [15]. We isolate the call-signs from the surveillance–broadcast (ADS-B) data fetched for each transcript and use them as context information. On average a sample contains 26 (Malorca), respectively 30 (LiveATC) surveillance call-signs.

The call-signs start generally with an airline identifier[5] (`lufthansa` ↔ `DLH`) followed by an alphanumeric call-sign number (`eight three kilo` ↔ `83K`). The call-sign number in the transcript is converted to its ICAO equivalent by the use of the NATO phonetic alphabet[6].

4. DATA AUGMENTATION

Each airspace has distinct characteristics like the occurrence of regional airlines. Noise levels of the voice channel can also vary, resulting in different WERs of the transcripts. Ideally, a CRU system could be fine-tuned to each new airspace by training it on a database for this region. In reality, there exist only a handful of ATC databases [4]. But not all of them contain labeled call-signs. A timestamp, as well as the location of the recordings are also missing in most cases, which makes it impossible to retrieve the corresponding surveillance information from the OSN database.

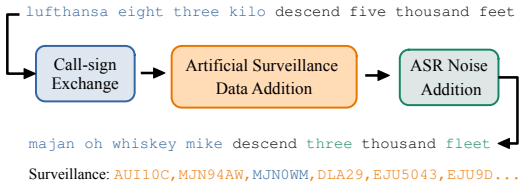

Fig. 1. Scheme of the data augmentation pipeline.

To overcome this issue, we propose a data augmentation pipeline which is shown in Figure 1. The basis for the pipeline is the Airbus training dataset [14], which contains roughly 28.000 transcripts with labeled call-signs. In the first step of the augmentation, the call-sign (`lufthansa eight three kilo`) is cut out of the transcript and replaced with an artificially generated call-sign (`majan oh whiskey mike`).

The rule-based data augmentation also includes real-life variations from the standard format. This includes missing identifiers, shortened call-sign numbers and the usage of different identifier formats. Transcript equivalents of `DLH72K` are for example `lufthansa seven two kilo`, `seven two kilo`, `lufthansa`, `lufthansa seventy-two kilo` and `dlh seven two kilo`.

In the next step, surveillance call-signs are added with the same parameters (number of call-signs with the same identifier, number of total call-signs, surveillance length) as real surveillance. To match the noise level of the test datasets (Malorca and LiveATC), simulated ASR noise (noisy distribution extracted from noisy ASR output) is introduced in the last step for the two noisy datasets (WER 7.0 and WER 30.0) but not for clean dataset (WER 0).

5. CONTEXT INTEGRATION

Context integration is necessary, since not all of the information loss through the ASR system can be recovered. If for example `five` would be missing in `Ryanair eight` **`five`** `three kilo`, the remaining `Ryanair eight three kilo` would be the wrong, but valid call-sign. A conventional CRU system would therefore predict `RYR83K` as target call-sign instead of `RYR853K`. Adding surveillance call-signs as additional input, as shown in Figure 2, allows the model to compensate the missing information in the transcript. With the timestamp of the transcript and its recording

[3]LiveATC Homepage: `https://www.liveatc.net/`
[4]OSN Homepage: `https://opensky-network.org/`
[5]A list of identifiers can be found here: `https://en.wikipedia.org/wiki/List_of_airline_codes`
[6]NATO phonetic alphabet: `https://en.wikipedia.org/wiki/NATO_phonetic_alphabet`

location, surveillance information can be retrieved from OSN (dotted path in Figure 2), like described in section 3.

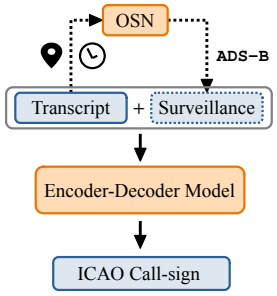

Fig. 2. The CRU system. The dotted path marks the optional surveillance retrieval via OSN with the aid of the transcripts timestamp and VHF receiver location.

6. EXPERIMENTAL SETUP

The basis of our CRU model is the `EncoderDecoderModel` from the Hugging Face transformers library [16], which showed superior performance over other designs in prior experiments. The language model head of this architecture allows also to predict call-signs without surveillance information. For both, encoder and decoder the pretrained `bert-base-uncased` model is used, to make use of the beneficial effect of using pretrained models for sequence-to-sequence tasks [17].

Table 2. Accuracy on the LiveATC testsets. The call-sign recognition models are trained on the augmented Airbus dataset with different WERs. Underlined accuracy scores symbolize the best vanilla recognition model, while bold scores mark the best model overall.

Taining sets	Accuracy on LiveATC test sets							
	WER 0		WER 28.4		WER 28.9		WER 33.1	
	Van	Sur	Van	Sur	Van	Sur	Van	Sur
WER 0	39.8	**89.4**	31.0	74.0	27.8	70.3	11.2	45.2
WER 7	60.2	88.6	47.2	**78.4**	44.0	**73.3**	16.4	**57.0**
WER 30	56.0	85.6	46.6	73.6	42.8	68.5	15.4	47.4

Since ATC speech transcripts differ highly from standard text, a domain adaptation is done by pretraining BERT (`bert-base-uncased`) on ATC transcripts using masked language modeling. The training of the CRU models is done on the augmented Airbus datasets. For each augmented dataset listed in Table 1 (WER 0, WER 7 and WER 30) a split of 40k/10k/10k for train/val/test sets is used. The models are either trained with (Sur) or without (Van) surveillance information. The transcript and the surveillance call-signs are concatenated and embedded into a single vector. This single or cross-encoder design allows interactions between the transcript and context from lower layers of the model on. The overall architecture for the Sur and Van models is the same, to ensure a fair comparison. The trained models are tested on the LiveATC and Malorca test sets listed in Table 1. The performance of all models is measured as accuracy or call-sign accuracy (CSA).

Table 3. Accuracy on the Malorca testsets. The call-sign recognition models are trained on the augmented Airbus dataset with different WERs. Underlined accuracy scores symbolize the best vanilla recognition model, while bold scores mark the best model overall.

Taining sets	Accuracy on Malorca test sets							
	WER 0		WER 6.42		WER 7.27		WER 8.47	
	Van	Sur	Van	Sur	Van	Sur	Van	Sur
WER 0	49.5	85.6	50.6	82.4	47.4	79.5	44.2	75.6
WER 7	53.8	**87.5**	53.6	84.9	50.4	83.5	46.8	80.7
WER 30	54.8	87.3	54.7	**85.0**	50.9	**83.7**	47.2	**81.0**

7. EXPERIMENTAL RESULTS

7.1. Surveillance Incorporation

Feeding the model surveillance call-signs not only allows recovering noisy ASR transcripts, that are lacking e.g. the airline identifier. The surveillance allows the model also to predict call-signs containing airline identifiers, that did not appear in the training data. Additionally the surveillance call-signs decrease the target space for the model.

Table 2 and Table 3 show the comparison between CRU models incorporating surveillance (Sur) and not incorporating surveillance (Van). The models that include surveillance call-signs outperform the vanilla models on every test set. On the high noise transcripts of the LiveATC dataset (WER 33.1), the benefit of the additional information shows the best. The vanilla network is here outperformed by a factor of 3-4. As an example of the recovering capabilities for noisy call-signs, we are able to predict 57% of the ICAO call-signs from the LiveATC transcripts (WER 33.1). Although they contain only 27% correct call-signs. This means an increase of 30%.

7.2. Noisy Training Data

To give the models more robustness against ASR noise, they are trained on different WER-level training data. On the Malorca data, the models trained on noisy transcripts (WER 7 and WER 30) outperform the model trained on clean transcripts (WER 0) on every test set. Both, the surveillance and the vanilla models benefit on similar levels from the training on noisy data, while the highest performance boost is reached

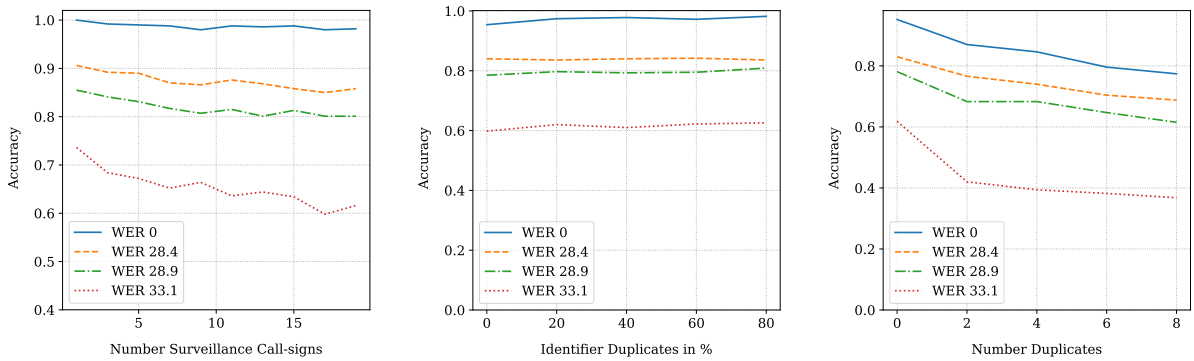

Fig. 3. Change of accuracy depending on (left) the number of call-signs in the surveillance data; (middle) the relative number of additional call-signs in the surveillance information containing the same call-sign identifier as the target call-sign; (right) the number of additional call-signs in the surveillance information containing the same call-sign number as the the target call-sign.

on the noisiest test set (WER 8.47) from 75.6% accuracy to 81.0%. On the noisier LiveATC test sets the overall mean accuracy of the vanilla model trained on WER 7 data is around 1.5 times higher than the accuracy of the model trained on noise-free data. Raising the WER of the training data further from 7 to 30 leads only to a small improvement on the high WER Malorca test sets.

The results show the benefit of training the model on (simulated) noisy transcripts if the target input of the model is ASR recognizer output. But more importantly, they also show, that even if there is just clean data available for training, including the surveillance call-signs is a necessary condition to reach maximum performance.

7.3. Surveillance Fluctuation Robustness

We investigate the robustness of our model against the three main surveillance parameters: number of call-signs in the surveillance, number of airline identifier duplicates and number of call-sign number duplicates. The evaluations are done on the LiveATC dataset by altering the surveillance information. The model trained on the WER 7 dataset is used for these tests since it performs the best on the noisy LiveATC test sets.

A higher number of surveillance call-signs increases the search space for the model. By increasing the surveillance size from 1 to 19, the accuracy decreases by 5% on the WER 28.4 test data, while there is a decrease of 12% on the WER 33.1 test set as Figure 3 shows. Intuitively this is clear since on noisy data, the model has to rely more on the additional context information.

Several airplanes of the same airline can be in the same airspace resulting in call-signs with an identical identifier (e.g `DLH124, DLH9M, DLH69F`). For the LiveATC and Malorca test set, each identifier in the surveillance occurs 1.45 respectively 1.9 times. Figure 3 shows, that the recognizer is very robust against airline identifier duplicates. Even with 80% of the surveillance call-sign identifiers being identical to the target identifier call-sign, there is no drop in accuracy.

In contrast to identifier duplicates, having the same call-sign number in the surveillance information (e.g. `DLH83K, CSA83K, RYR83K`) is quite rare. In the LiveATC dataset, only in 2.7% of the cases, a call-sign number appears twice in the surveillance information. With one duplicate of the target call-sign, which is already higher then what can be expected in the real-life scenario, the accuracy drop on the WER 28.4 and WER 28.9 dataset stays below 5% as Figure 3 shows. For the high noise dataset, the drop is around 10%.

8. CONCLUSION

In this work, we have introduced a method for enhancing call-sign recognition and understanding (CRU) by incorporating context information in the form of surveillance call-signs without changing the model architecture. We have shown that this improves the call-sign accuracy up to 4 times. Our data augmentation pipeline allows to generate training data for specific airspaces, even if there are no transcripts available for that region. We have shown that introducing ASR noise in the data augmentation pipeline improves the vanilla model performance up to 1.5 times.

We can show that our models are robust against the occurrence of multiple surveillance call-signs containing the same identifier. The number of included surveillance call-signs included should be kept as low as possible since the call-sign accuracy decreases linearly with the number of the surveillance call-signs. For the rare case of an additional call-sign occurring with the target call-sign number, we can show that the accuracy drop stays below 5% for the low call-sign WER test sets and under 10% for the high WER call-sign test set.

In the future, we want to look also at other context incorporation methods. We also plan to adapt our model to other named entities appearing in ATC transcripts like commands and values.

9. REFERENCES

[1] S. Eskilsson, H. Gustafsson, S. Khan, and A. Gurtov, "Demonstrating ADS-B and CPDLC Attacks with Software-Defined Radio," *Integrated Communications, Navigation and Surveillance Conference, ICNS*, vol. 2020-Septe, pp. 1–9, sep 2020.

[2] M. Kleinert, H. Helmke, G. Siol, H. Ehr, A. Cerna, C. Kern, D. Klakow, P. Motlicek, Y. Oualil, M. Singh, and A. Srinivasamurthy, "Semi-supervised adaptation of assistant based speech recognition models for different approach areas," in *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. dec 2018, vol. 2018-Septe, Institute of Electrical and Electronics Engineers Inc.

[3] A. Srinivasamurthy, P. Motlicek, M. Singh, Y. Oualil, M. Kleinert, H. Ehr, and H. Helmke, "Iterative learning of speech recognition models for air traffic control," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. September, pp. 3519–3523, 2018.

[4] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Vesely, and R. Braun, "Automatic speech recognition benchmark for air-traffic communications," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, jun 2020, vol. 2020-Octob, pp. 2297–2301.

[5] T. Shore, F. Faubel, H. Helmke, and D. Klakow, "Knowledge-based word lattice rescoring in a dynamic context," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, vol. 2, pp. 1082–1085.

[6] A. Schmidt, Y. Oualil, O. Ohneiser, M. Kleinert, M. Schulder, A. Khan, H. Helmke, and D. Klakow, "Context-based recognition network adaptation for improving on-line ASR in air traffic control," in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, pp. 13–15.

[7] Y. Oualil, M. Schulder, H. Helmke, A. Schmidt, and D. Klakow, "Real-time integration of dynamic context information for improving automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, vol. 2015-Janua, pp. 2107–2111.

[8] M. Kocour, K. Veselý, A. Blatt, J. Zuluaga Gomez, I. Szöke, J. Černocký, D. Klakow, and P. Motlicek, "Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition," pp. 3301–3305, 2021.

[9] I. Nigmatulina, R. Braun, J. Zuluaga-Gomez, and P. Motlicek, "Improving callsign recognition with air-surveillance data in air-traffic communication," aug 2021.

[10] D. Povey, G. Boulianne, L. Burget, P. Motlicek, and P. Schwarz, "The Kaldi Speech Recognition," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, , no. January, 2011.

[11] 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*, oct 2019, vol. 2019-Septe, pp. 2993–2997.

[12] V. Gupta, L. Rebout, G. Boulianne, P. Ménard, and J. Alam, "CRIM's Speech Transcription and Call Sign Detection System for the ATC Airbus Challenge Task," in *Interspeech 2019*. sep 2019, pp. 3018–3022, International Speech Communication Association.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186.

[14] 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 *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, pp. 2866–2870.

[15] M. Schäfer, M. Strohmeier, V. Lenders, I Martinovic, and M. Wilhelm, "Bringing up OpenSky: A large-scale ADS-B sensor network for research," in *IPSN 2014 - Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (Part of CPS Week)*. 2014, pp. 83–94, IEEE Computer Society.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, and C Xu, "Transformers: State-of-the-Art Natural Language Processing," 2020, pp. 38–45.

[17] S. Rothe, S. Narayan, and A. Severyn, "Leveraging Pre-trained Checkpoints for Sequence Generation Tasks," Tech. Rep., 2020.