A TWO-STEP APPROACH TO LEVERAGE CONTEXTUAL DATA: SPEECH RECOGNITION IN AIR-TRAFFIC COMMUNICATIONS

Iuliia Nigmatulina [†,‡], *Juan Zuluaga-Gomez* [†,§], *Amrutha Prasad* [†,¶], *Seyyed Saeed Sarfjoo* [†], *Petr Motlicek* [†]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] Institute of Computational Linguistics, University of Zürich, Switzerland
[§] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[¶] Brno University of Technology, Brno, Czech Republic

ABSTRACT

Automatic Speech Recognition (ASR), as the assistance of speech communication between pilots and air-traffic controllers, can significantly reduce the complexity of the task and increase the reliability of transmitted information. ASR application can lead to a lower number of incidents caused by misunderstanding and improve air traffic management (ATM) efficiency. Evidently, high accuracy predictions, especially, of key information, i.e., callsigns and commands, are required to minimize the risk of errors. We prove that combining the benefits of ASR and Natural Language Processing (NLP) methods to make use of surveillance data (i.e. additional modality) helps to considerably improve the recognition of callsigns (named entity). In this paper, we investigate a two-step callsign boosting approach: (1) at the 1st step (ASR), weights of probable callsign n-grams are reduced in G.fst and/or in the decoding FST (lattices), (2) at the 2nd step (NLP), callsigns extracted from the improved recognition outputs with Named Entity Recognition (NER) are correlated with the surveillance data to select the most suitable one. Boosting callsign n-grams with the combination of ASR and NLP methods eventually leads up to 53.7% of an absolute, or 60.4% of a relative, improvement in callsign recognition.

Index Terms— automatic speech recognition, human-computer interaction, Air-Traffic Control, Air-Surveillance Data, Callsign Detection, finite-state transducers

1. INTRODUCTION

Key components of speech communication between pilots and Air-Traffic Controllers (ATCo), i.e., callsigns, which are used for identification of aircrafts, and providing commands, demand high recognition accuracies. Callsigns are unique identifiers for aircrafts, of which the first part is an abbreviation of airline name and the last part is a flight number that contains a digit combination and may also incorporate an additional character combination, e.g., *TVS84J* (see Table 1). At a certain time point, only few aircrafts are usually in the radar zone which means only a limited number of callsigns can be referred to in the ATCo communications. If a recognized callsign does not match any 'active' callsign registered by radar at the given time point, it means that there is no corresponding aircraft in the air space and the automatically recognized command (from voice communication) is invalid. Therefore, contextual information coming from the surveillance (radar) data allows adjusting system predictions that can significantly increase its accuracy.

Table 1. Callsigns: compressed and extended (airlines designators are in bold)

Callsign	Extended callsign
SWR2689	**swiss** two six eight nine
RYR1RK	**ryanair** one romeo kilo
RYR1SG	**ryanair** one sierra golf

Although contextual information has been already used in previous ATC studies [1–4], or more recently in [5–7]; it has been never adapted for both ASR and concept extraction outputs simultaneously and without a need of any additional knowledge (e.g., manual annotation, classes, etc.). This research aims to leverage the available contextual information by combining ASR and NLP methods. We believe that ASR and NLP are complementary tasks rather than separated ones. Whereas ASR exploits speech to produce a sequence of words, NLP exploits the intrinsic characteristics in a given snippet of text. ASR normally struggles to model long sequences, while state-of-the-art NLP systems allow extracting key information in the whole chunks of text; for instance an entire ATC utterance. In the proposed approach, we focus on an iterative use of contextual data, to take advantage of a combination of ASR and NLP modules. (1) First, boosting the probability of active callsigns in ASR system (*FST-boosting*), (2) second, boosting ASR outputs (*NLP-boosting*) in order to correct those predicted callsigns, which are not present in the surveillance data.

The rest of the paper is organised as follows: Section 2 reviews current approaches on integrating contextual knowledge in ASR for ATC communications. Section 3 gives a theoretical background of the proposed ASR-NLP approach to leverage surveillance data. Then, we present the data and the experiment set up in Section 4. Finally, we report the results and summarise our observations and ideas in Section 5 and 6, respectively.

2. CONTEXTUAL INFORMATION FOR CALLSIGN DETECTION

Contextual data on the ASR level can be integrated by modifying weights of target n-grams in the grammar or/and in the ASR decoding lattices, e.g. by mean of generalised composition of baseline

The work was supported by the European Union's Horizon 2020 projects. No. 864702 - ATCO2 (Automatic collection and processing of voice data from air-traffic communications), which is a part of Clean Sky Joint Undertaking. The work was partially supported by SESAR Joint Undertaking under Grant Agreement No. 884287 - HAAWAII (Highly automated air traffic controller workstations with artificial intelligence integration).

LM and Weighted Finite State Transducers (WFSTs) with the target contextual n-grams [8–10]. A similar approach has been recently adopted in the ATC domain [5, 6] and proved to offer a significant gain in callsign recognition. A list of callsigns to be boosted is regularly changing and needs to be updated dynamically per each utterance. Thus, weights of callsign n-grams are dynamically modified in the WFST. The first of the methods is lattice rescoring, where the weights are adjusted on the word recognition lattices from the first pass decoding. In the other method, weights are dynamically modified directly in the grammar (G.fst), which allows having target n-grams boosted before the decoding is performed [6]. For our experiments, we will adopt the lattice rescoring approach to leverage the performance on the ASR side.

Besides the ASR performance, contextual information for ATC has been also used to improve concept extraction [1–4]. Schmidt et al. [1] applied a Context-Free Grammar (CFG)-based LM limiting the search space according to the contextual data. Shore et al. [2] and Oualil et al. [3, 4] build a CFG-based concept extractor with all semantic concepts of ATC embedded in XML annotation tags. In [2], after decoding, the lattice hypotheses are rescored by incorporating an additional knowledge source component to the cost function. The knowledge-based rescoring penalises hypotheses that are invalid in the context, e.g., callsigns not registered in the air space. In [3], to overcome the problem of variability of ATCO commands, the weighted Levenshtein distance is applied to find the closest match between an ASR hypothesis and generated context word sequences. [4] combines methods from [2, 3] adding more contextual constraints from data with temporal information. Although these methods help to considerably increase the recognition accuracy, their limitation is that it deals only with concepts and callsigns which are annotated and included into the grammar. Those n-grams that do not appear in the grammar can not be extracted and evaluated. Finally, Helmke et al. [11] recently proposed a machine learning algorithm for command extraction from the ASR hypothesized outputs with the use of keywords. This model achieves good results and it is the second alternative approach to our methods.

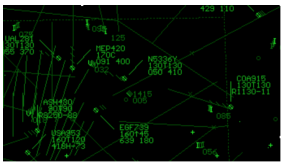
3. METHODS

We focus on the combination of ASR and NLP methods and investigate two-steps approach for callsigns extraction. As a callsign is a sequence of words, using contextual information to improve recognition of callsigns is a task of boosting n-grams. The contextual data comes from radar in a compressed form, i.e., standardized phraseology format of International Civil Aviation Organization (ICAO) [12] (see Fig. 1). To introduce the contextual knowledge into the ASR system, all callsigns need to be expanded to word sequences (Table 1). The compressed form often allows more than one possible realisation in the ATCos' speech: For example, **DLH5KX** can be expanded as *'hansa five kilo x-ray'* or *'lufthansa five kilo x-ray'*, etc. As we can not say which particular expansion is true for an uttered callsign, it is important to take all expansion variants into account.

3.1. Integration of contextual knowledge into ASR system

In a standard hybrid-based ASR system, the different knowledge sources are represented as WFSTs, which are combined by the 'composition' operator together in the final decoding graph [13]. Information from additional knowledge sources can be also integrated into a system by means of composition.

Our first integration of contextual knowledge into ASR is done on the LM level (*G-extension*). The idea is to boost callsign n-grams



AL78 ACA854 AFR6735 BEE6BU CLX79W CLX98G DAL32 DLH6LY EIN23W EXS155 EXS95LP EXS9DL EZY61MT EZY82NY HMD61 ICE416 RYR1RK RYR3TA RYR4DD RYR4TM RYR50HD RYR7245 SAS99K STK19L STK21GL STK21M STK31S STK3212

Fig. 1. Callsigns in ICAO format received from radar.

already available in LM, and even more important to add those callsign n-grams, which are absent (e.g., >3 words sequences in 3-gram LM). We build a contextual *FST* that includes all possible callsigns from the tower: all callsigns registered by the radar at different time stamps (from 17K to 280K callsigns to boost in different test sets; see last column in Table 2). Then, the main $G.fst$ is composed with the contextual $G\_biased.fst$ and the result of composition is used in the final decoding $HCLG$ graph.

The second integration of contextual information (*lattice rescoring*) is done per utterance on top of the decoding lattices which allows flexible adaptation to new-coming contextual information avoiding changing the main decoding graph ($HCLG$) (for more details check [6]). Weights in lattices are rescored according to the surveillance data: for each test utterance, an FST biased to callsigns n-grams registered at the time stamp when an utterance is created and composed with lattices created in the first pass:

$$Lattices' = Lattices \circ biasing\_FST \quad (1)$$

Weights updated in the composition are used for final predictions.

3.2. Integration of contextual knowledge on ASR transcripts

Our approach for integrating contextual knowledge on ASR transcripts (e.g., 1-best hypothesis) is based on a two-step pipeline. Each step conveys an independent module.

3.2.1. Named Entity Recognition (NER) module

ATC communications carry rich information such as callsigns, commands, values and units; they can be seen as 'named entities'. We propose a NLP-based system to extract such information from ASR transcripts. We defined callsigns, commands, units, values, greetings OR the rest (e.g., 'None' class) as tags for the NER task, as depicted in Figure 2. First, we downloaded a BERT [14] model pre-trained as masked language model from Huggingface [15] and fine-tuned it on NER task with 12k sentences (∼12 hours of speech), where each word has a tag. Then, we developed a data augmentation pipeline in order to increase the amount of training data: 1M samples from 12k sentences. The pipeline has four actions that modifies the training sample: *add*, *delete*, *swap*, or *move* the **callsign** across the utterance -sentence-. *Delete* and *move* actions, remove and keep the same callsigns, respectively; *add* and *swap* generate a sentence with a new callsign picked randomly from a callsign list. The callsign list is pre-defined by a user, which makes the approach easy to deploy in out-of-domain data (i.e., callsigns from different airports/countries).

3.2.2. Re-ranking module based on Levenshtein distance

The BERT-based system for NER allows us to extract the callsign from a given transcript or ASR 1-best hypotheses. Recognition of this entity is crucial where a single error produced by the ASR system affects the whole entity (normally composed of three to eight

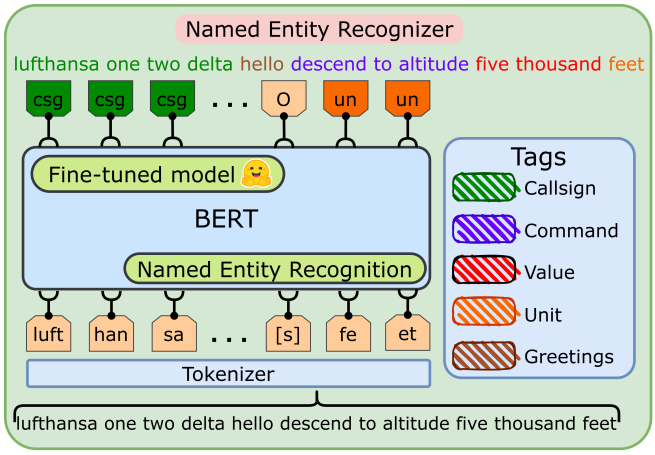


Fig. 2. BERT-based model (Huggingface) fine-tuned on NER task.

words). Additionally, speakers regularly shorten callsigns in the conversation making it impossible for an ASR system to generate the full entity (e.g., *'three nine two papa'* instead of *'austrian three nine two papa'*, *'six lima yankee'* instead of *'hansa six lima yankee'*). One way to overcome this issue is to re-rank entities extracted by the BERT-based NER system with the surveillance data. The output of an NER system is a list of tags that match words or sequences of words in an input utterance. As our only available source of contextual knowledge are callsigns registered at a certain time and location, we extract callsigns with the NER system and discard other entities. Correspondingly, each utterance has a list of callsigns expanded into word sequences (shown in Table 1). As input, the re-ranking module takes (i) a callsign extracted by the NER system and (ii) an expanded list of callsigns. The re-ranking module compares a given n-gram sequence against a list of possible n-grams, and finds the closest match from the list of surveillance data based on the weighted Levenshtein distance. We skip the re-ranking in case the NER system outputs a 'NO_CALLSIGN' flag (no callsign recognized).

4. DATA AND EXPERIMENTAL SETUP

4.1. Data

For the callsign boosting experiments, we use four test sets; all of them have utterances both with and without callsigns (see Table 2).

LiveATC: the first test set is from the LiveATC[1] data recorded from publicly accessible VHF radio channels, which includes both pilots and ATCo speech and, therefore, is of rather low quality (i.e., low SNR often below 10dB) [16].

MALORCA: Prague and Vienna test sets are mainly of good quality (i.e., telephone quality speech with SNR usually above 20dB) data from the MALORCA project [17, 18][2] which includes only ATCo speech. The recognition accuracy of the baseline model are already high above the one reached on VHF LiveATC data (see Table 3). The data was collected from the Prague and Vienna airports and, thus, forms two separate sets correspondingly.

NATS: a data set collected under HAAWAII project[3] with the data coming from London approach (airport). This data is relatively high-quality, similar to MALORCA.

The data sets are used differently in training ASR and NER models. The ASR train data includes Malorca sets but not LiveATC and NATS. The data for fine-tuning the NER system contains LiveATC data but neither Malorca, nor NATS sets.

[1]Streaming audio platform that gathers VHF aircraft communications

[2]From the 'standard' MALORCA test sets [18] only utterances with the available surveillance information are selected.

[3]https://www.haawaii.de/wp/

Table 2. Test sets (callsigns (csgn) per utterance (utt) — median of callsigns per utterance in the surveillance data)

Test set	N of utt with a csgn	N of utt w/o a csgn	Csgn per utt	Min	All csgns
LiveATC	581	29	28	40	280K
Malorca Prague	784	88	5	82	17K
Malorca Vienna	877	38	19	65	59K
NATS	794	73	50	50	168K

4.2. ASR model

For training the baseline acoustic model, as well as for the decoding and rescoring experiments, we used the Kaldi framework [19]. The system follows the standard Kaldi recipe, which uses MFCC and i-vectors features. The standard chain training is based on Lattice-free MMI (LF-MMI) [20], which includes 3-fold speed perturbation and one third frame sub-sampling.

The acoustic model is a CNN-TDNNF trained on approximately 1200 hours of ATC labeled augmented data [16, 21]. First, the training databases (195 hours[4]) were augmented by adding noises that match LiveATC audio channel (one batch between 5-10 dB and other 10-20dB SNR). Afterwards, we applied speed perturbation, obtaining almost 1200 hours of training data. The model was further improved with 700 hours of semi-supervised data collected in LiveATC for different airports from Europe [17]. The LM is 3-gram trained on the same data as the acoustic model with an additional textual data from additional public resources such as airlines names, airports, ICAO alphabet and way-points in Europe.

4.3. Evaluation

Since this paper focuses on improving callsign detection, we evaluate the proposed methods by calculating the accuracy of callsign extraction. For the evaluation we use ICAO format, which is the target form to display on the screen of ATCo and pilots, and we have only two outcomes: ICAO is recognized 'correctly' VS 'incorrectly'. In the previous studies [5, 6], the accuracy of callsign recognition is evaluated with matching the ground truth callsign n-grams to the ones in utterances. This approach, however, does not correspond to the real situation, when ground truth callsigns are not available. In our experiments, we do not only do speech recognition but proceed with callsign extraction, we evaluate the performance directly on the extracted entities. In addition, the use of the ICAO format helps to avoid issues with variability of pronunciation within a callsign: the full form of callsign is extracted automatically but a speaker says a shorten version, which is then outputted by the ASR, as well as recorded in the ground truth transcriptions (see example above 3.2.2). All experiments share the same ASR and BERT-based NER systems, as well as the ICAO extractor module; thus, the performances are only impacted by the proposed boosting techniques.

[4]The ATCO2 test set is publicly available in https://www.atco2.org/data

Table 3. Results of callsign extraction with ASR boosting (ASR-B) and post-boosting (NLP-B): the accuracy of callsign recognition (%) is calculated for the callsigns in ICAO format (see Section 4.3)

Method				**Test sets (callsign recognition accuracy)**			
				LiveATC	Prague	Vienna	NATS
ASR output	Callsign extraction (baseline)			42.8	64.4	48.4	35.2
	Lattice rescoring	G-extension	NLP-boosting				
	✓	-	-	53.1	66.9	59.6	37.1
	-	✓	-	44.4	64.3	49.2	34.8
	✓	✓	-	52.8	66.9	52.1	36.8
	-	-	✓	88.4	**95.0**	**86.0**	87.0
	✓	-	✓	**88.5**	94.8	84.3	**88.9**
	-	✓	✓	87.7	**95.0**	85.6	88.2
	✓	✓	✓	88.0	94.7	84.0	88.0
Ground Truth	Callsign extraction (oracle)			**89.7**	72.2	59.6	67.4
	+ NLP-Boosting			89.3	**95.4**	**87.0**	**94.0**
ASR WER	**(without boosting)**			32.4	3.4	9.2	24.4

5. RESULTS

As a baseline we use callsign extraction done directly on the outputs of our ASR system. Then, we apply the proposed boosting techniques (G-extension, lattice rescoring, NLP-boosting) in different combinations to see how they can benefit from each other. In Table 3, the results of the experiments are presented on four different test sets with accuracy of callsign (ICAO) recognition. Overall, the proposed metrics help to improve the baseline accuracy from 30.6% to 53.7% absolutely, or from 32.1% to 60.4% relatively (for the test sets Prague and NATS correspondingly; when the NATS set gets the highest improvement being the out-of-domain data). The best results are always achieved with the use of NLP-boosting. For LiveATC and NATS sets, the out-of-domain sets in the ASR training, the best performance is achieved with the combination of NLP-boosting and ASR-boosting (lattice rescoring) methods.

At the same time, the G-extension has a contradicting effect. It helps to improve results comparing to the baseline for the LiveATC and Vienna sets, yet, its combination with lattice rescoring achieves worse accuracy than lattice rescoring alone. The possible drawback of the G-extension method is that a very high number of available callsigns are boosted in LM FST (see last column 2). It can introduce confusion when combining with the lattice rescoring boosting method, which focuses on only current callsigns. On the other hand, it does not need any modifications during the decoding and serves as a general domain adaptation. Thus, G-extension can be used to improve the outputs when other methods are not available, otherwise, can be skipped. The number of callsigns used to boost the ASR outputs may also have the degradation effect on the performance of the lattice rescoring approach. Although in this case, the number of callsigns did not exceed 50, we investigated its impact. The test sets have different numbers of boosted n-grams, from 5 to 50 (see Table 1), but even with 50 boosted callsigns the recognition accuracy goes considerably up comparing to the baseline.

Along with the evaluation of boosting methods on the ASR outputs, we provide the 'oracle' results, when callsigns are extracted on the ground truth transcriptions (2nd line in Table 3). This comparison allows estimating the impact of the proposed methods to the callsign extraction improvement, when no ground truth information is available. Even if the 'oracle' scores always stay better, the accuracy achieved with our systems shows close and comparable results. No improvement with NLP-boosting on the ground truth transcription for LiveATC test set can be explained by already high accuracy of callsign extraction, as LiveATC data was used to fine-tune the NER.

Table 4. Examples of improved callsign recognition (bold part)

Baseline (incorrect ICAO)	Boosted (correct ICAO)
wizz air four one six (**WZZ**416)	**iceair** four one six (**ICE**416)
easy three delta (**EZY**3D)	**fraction eight eight** three delta (**NJE88**3D)
serbia one nine lima (**ASL**19L)	**stobart** one nine lima (**STK**19L)

Table 4 gives examples of improvement where airline names and callsigns are detected correctly comparing to the baseline predictions. Our methods demonstrate consistent results for data of different quality. The level of noise in the recordings of LiveATC and Malorca test sets is very different, as well as WERs achieved by their baseline ASR systems (the last line in Table 3; [6]). Nevertheless, we see considerable improvement for all test sets and the general tendency stays the same. The main advantage of the proposed approach comparing to the others is its simplicity and flexibility. The NER-system can be fine-tuned to different data sets that makes it easy to adapt to new out-of-domain data. Moreover, it is also suitable for the online implementation.

6. CONCLUSION

We investigated a two-step approach of integrating contextual radar data in order to dynamically improve the recognition of callsigns per utterance. We demonstrated that the best result is achieved with the NLP-boosting and with the combination of NLP-boosting and lattice rescoring methods on all test sets of different recording quality with the significant improvement, i.e., from 32.1% to 60.4% of relative improvement on callsign recognition accuracy across the evaluated data sets. Introduction of contextual information considerably improves recognition of callsigns and, thus, recognition of ATCo messages in general. As a noisy environment leading to lower recognition accuracy is often a reality in pilot-ATCo communication, the proposed methods and their combination will definitely benefit the recognition of the key information in ATCo speech.

7. REFERENCES

[1] Anna Schmidt, Youssef Oualil, Oliver Ohneiser, Matthias Kleinert, Marc Schulder, Arif Khan, Hartmut Helmke, and Dietrich Klakow, "Context-based recognition network adaptation for improving on-line asr in air traffic control," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 13–18.

[2] Todd Shore, Friedrich Faubel, Hartmut Helmke, and Dietrich Klakow, "Knowledge-based word lattice rescoring in a dynamic context," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] Youssef Oualil, Marc Schulder, Hartmut Helmke, Anna Schmidt, and Dietrich Klakow, "Real-time integration of dynamic context information for improving automatic speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] Youssef Oualil, Dietrich Klakow, Gyorgy Szaszák, Ajay Srinivasamurthy, Hartmut Helmke, and Petr Motlicek, "A context-aware speech recognition and understanding system for air traffic control domain," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 404–408.

[5] Martin Kocour, Karel Veselỳ, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, Jan Cernocky, Dietrich Klakow, and Petr Motlicek, "Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3301–3305.

[6] Iuliia Nigmatulina, Rudolf Braun, Juan Zuluaga-Gomez, and Petr Motlicek, "Improving callsign recognition with air-surveillance data in air-traffic communication," Idiap Research Institute, 2021, pp. 1–5, Idiap Research Institute.

[7] Juan Zuluaga-Gomez, Iuliia Nigmatulina, Amrutha Prasad, Petr Motlicek, Karel Veselỳ, Martin Kocour, and Igor Szöke, "Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems," in *Proc. Interspeech 2021*, 2021, pp. 3296–3300.

[8] Keith Hall, Eunjoon Cho, Cyril Allauzen, Francoise Beaufays, Noah Coccaro, Kaisuke Nakajima, Michael Riley, Brian Roark, David Rybach, and Linda Zhang, "Composition-based on-the-fly rescoring for salient n-gram biasing," 2015.

[9] Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno, "Bringing contextual information to google speech recognition," 2015.

[10] Jack Serrino, Leonid Velikovich, Petar S Aleksic, and Cyril Allauzen, "Contextual recovery of out-of-lattice named entities in automatic speech recognition.," in *Interspeech*, 2019, pp. 3830–3834.

[11] Hartmut Helmke, Matthias Kleinert, Oliver Ohneiser, Heiko Ehr, and Shruthi Shetty, "Machine learning of air traffic controller command extraction models for speech recognition applications," in *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*. IEEE, 2020, pp. 1–9.

[12] "All clear phraseology manual," in *Eurocontrol, Brussels, Belgium*, 2011, "[Online; accessed 10-September-2021]".

[13] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz et al, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45, Association for Computational Linguistics.

[16] Juan Zuluaga-Gomez, Karel Veselỳ, Alexander Blatt, Petr Motlicek, Dietrich Klakow, Allan Tart, Igor Szöke, Amrutha Prasad, Saeed Sarfjoo, Pavel Kolčárek, et al., "Automatic call sign detection: Matching air surveillance data with air traffic spoken communications," in *Multidisciplinary Digital Publishing Institute Proceedings*, 2020, vol. 59, p. 14.

[17] Banriskhem Khonglah, Srikanth Madikeri, Subhadeep Dey, Hervé Bourlard, Petr Motlicek, and Jayadev Billa, "Incremental semi-supervised learning for multi-genre speech recognition," in *Proceedings of ICASSP 2020*, 2020.

[18] Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan, Gyorgy Szaszak, Youssef Oualil, and Hartmut 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.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[20] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016, pp. 2751–2755.

[21] Juan Zuluaga-Gomez, Petr Motlicek, Qingran Zhan, Karel Veselỳ, and Rudolf Braun, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," in *Proc. Interspeech 2020*, 2020, pp. 2297–2301.