Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition

Amrutha Prasad*[,1,2], Juan Zuluaga-Gomez[1,3], Petr Motlicek[1,2], Saeed Sarfjoo[1], Iuliia Nigmatulina[1,4], Oliver Ohneiser[5], Hartmut Helmke[5]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Brno University of Technology, Speech@FIT, IT4I CoE, Brno, Czech Republic
[3]Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland
[4]Institute of Computational Linguistics, University of Zurich, Switzerland
[5]German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany
*corresponding author: amrutha.prasad@idiap.ch

***Abstract*—Automatic Speech Recognition (ASR) for air traffic control is generally trained by pooling Air Traffic Controller (ATCO) and pilot data into one set. This is motivated by the fact that pilot's voice communications are more scarce than ATCOs. Due to this data imbalance and other reasons (e.g., varying acoustic conditions), the speech from ATCOs is usually recognized more accurately than from pilots. Automatically identifying the speaker roles is a challenging task, especially in the case of the noisy voice recordings collected using Very High Frequency (VHF) receivers or due to the unavailability of the push-to-talk (PTT) signal, i.e., both audio channels are mixed. In this work, we propose to (1) automatically segment the ATCO and pilot data based on an intuitive approach exploiting ASR transcripts and (2) subsequently consider an automatic recognition of ATCOs' and pilots' voice as two separate tasks. Our work is performed on VHF audio data with high noise levels, i.e., signal-to-noise (SNR) ratios below 15 dB, as this data is recognized to be helpful for various speech-based machine-learning tasks. Specifically, for the speaker role identification task, the module is represented by a simple yet efficient knowledge-based system exploiting a grammar defined by the International Civil Aviation Organization (ICAO). The system accepts text as the input, either manually verified annotations or automatically generated transcripts. The developed approach provides an average accuracy in speaker role identification of about 83%. Finally, we show that training an acoustic model for ASR tasks separately (i.e., separate models for ATCOs and pilots) or using a multitask approach is well suited for the noisy data and outperforms the traditional ASR system where all data is pooled together.**

***Keywords*—assistant based speech recognition, air traffic management, multitask acoustic modeling, speaker role classification, Kaldi**

I. Introduction

Previous work [1], [2] as part of the MALORCA[1] and AcListant-Strips[2] projects, focused on i) improving Assistant-Based Speech Recognition (ABSR) accuracy only for ATCOs, ii) reducing workload for ATCOs [3], and iii) increasing efficiency [4] of ATCOs. In the ongoing HAAWAII[3] project, research focuses on developing a reliable and adaptable solution to transcribe voice commands issued by both ATCOs and pilots automatically.

An error-resilient and accurate ASR system is critical in the ATC domain. Current state-of-the-art technologies require large amounts of data to train ASR systems. The goal of a recently finished project called ATCO2[4] was to collect and automatically transcribe a large database of voice recordings of ATCOs and pilots (with a minimum effort) for the purpose mentioned above [5]–[7]. ATCO and pilot speech recordings are usually pooled together [1], [6], [8] to train the ASR despite having a significant variability in the data distribution (acoustic and grammatical conditions), and the number of speakers in the data. As a result of the variability in the data distribution, ASR performances are significantly different for ATCO and pilot speech.[5] Our baseline system trained by pooling all data and evaluated on noisy data (signal-to-noise ratio (SNR) below 15 dB) shows an absolute difference in word error rate (WER) of 9.7% on ATCO and pilot speech (ATCO WER: 36.1%, Pilot WER: 45.8%). ASR on another dataset also revealed that it is 'twice as hard ' to correctly recognize pilot utterances compared to ATCO utterances due to shortened speech [9].

In this paper, we hypothesize that introducing information about the speaker role during ASR training can help mitigate this variability. We consider the approach of training the acoustic model in the ASR to produce outputs for each speaker role–ATCO and pilot–separately. This is often called multitask

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

[2]Active Listening Assistant Strips: https://www.malorca-project.de/wp/?page_id=350

[3]Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration: https://www.haawaii.de

[4]Automatic collection and processing of voice data from air-traffic communications https://www.atco2.org/.

[5]The air traffic controllers' speech is more straightforward to recognize than pilots'.

learning in Deep Neural Networks parlance [10]. Specifically, this paper investigates a multitask approach to training AMs to be integrated into ASR for ATCO and pilot. The obvious first step is automatically splitting the ATC speech communication data into two speaker roles. However, obtaining speaker labels manually on a large dataset is expensive and time-consuming. A common approach is to diarize the audio [11], [12] (see Section II-A). Although the ATCO speech is often cleaner (higher SNR value) than the pilot (as the former communicates from a controlled acoustic environment), the speech recordings collected in ATCO2 project using Very High Frequency (VHF) receivers[6] are noisy for both ATCO and pilot channels [13]. In such a case, a speaker diarization system may fail to assign speaker labels accurately. Thus, it is not advisable to rely fully on a pure acoustic-based system to obtain accurate speaker labels.

Another approach to obtain the speaker class is through leveraging the 'ICAO' grammar to classify an utterance as one of the classes based on text. The ICAO grammar [14] is a well-defined, standard phraseology to ensure safe communication between a controller and pilot, which in turns assure smooth travel of the aircraft. Once the speaker role labels are available over a large database, AMs can be trained for both ATCOs and pilots with different approaches. In this study, we show that due to the poor acoustic conditions, training a single acoustic model (AM) by pooling all data can result in degradation of the ASR performance on pilot speech. To obtain better accuracy, AM should be trained separately for ATCO and pilot data, or considered as different tasks by using a multitask approach.

Our paper is structured as follows. Section 2 provides a brief overview of the work related to multitask automatic speech recognition. The datasets used are described in Section 3 followed by Section 4 that describes speaker role classification with text. Section 5 explains the experimental setup and the results obtained, which are followed by the conclusion in Section 6.

II. Related Work

A. Speaker Role Classification

The current approach to identify a speaker role for a given utterance is extracted from an acoustic-based diarization system. A speaker diarization system can be defined as a task of defining speaker roles to segments of an utterance. This approach is developed over the years for performing segmentation of conversations or for finding a speaker of interest in a given set of speakers. Several approaches based on Bayesian hidden Markov model and neural networks have evolved over the past two decades, which are current state-of-the-art.

[6]Blog related to setting up receivers https://www.atco2.org/news/setting-up-vhf-receiver-for-air-traffic-communication

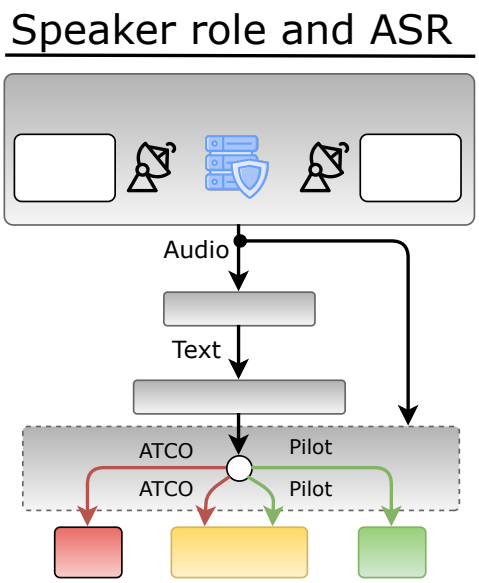

Figure 1. **Overall pipeline to train a multitask ABSR system**. A seed ABSR is used to transcribe all the available VHF data. Then we use the grammar-based module to split the databases by speaker roles, i.e., ATCO or pilot. Then, the split databases are used to train speaker-dependent acoustic models in the case of separate ATCO and pilot models. The same information is also used by the multitask system to select the task to be trained for each utterance. The same procedure is applied to other datasets used in this paper.

Communication between ATCOs and pilots is collected as a single channel, typically without VAD; thus, the obtained audio files end spamming several minutes. Therefore, the first step conveys to apply segmentation, i.e., typically a Voice Activity Detection (VAD) system. This system splits the audio based on long silence regions to get a set of audio files to which diarization can then be applied. The current acoustic-based speaker role classification system used in HAAWAII employs the clustering of speaker embeddings (x-vectors) approach described by Landini et al. [15]. Firstly, a neural network is trained to discriminate between various speakers, so embeddings can capture the relevant information, which allows comparing speech fragments and deciding if they belong to the same speaker. The output of this neural network is the x-vectors. The second step is clustering these x-vectors based on the Bayesian hidden Markov model, where each state in the model is represented as one speaker. When finding the state that most likely is produced by a given x-vector, the x-vector is assigned one speaker. The final diarization output is the assignment of an x-vector to a speaker role.

In a typical recording of ATC communications, one ATCO would communicate with several pilots. Hence, most parts of these speeches are marked as ATCOs and the rest as pilots based on the former prior assumption. Since our primary goal is to identify ATCO/pilot, many speakers share the same speaker role label, and thus a speaker diarization system will

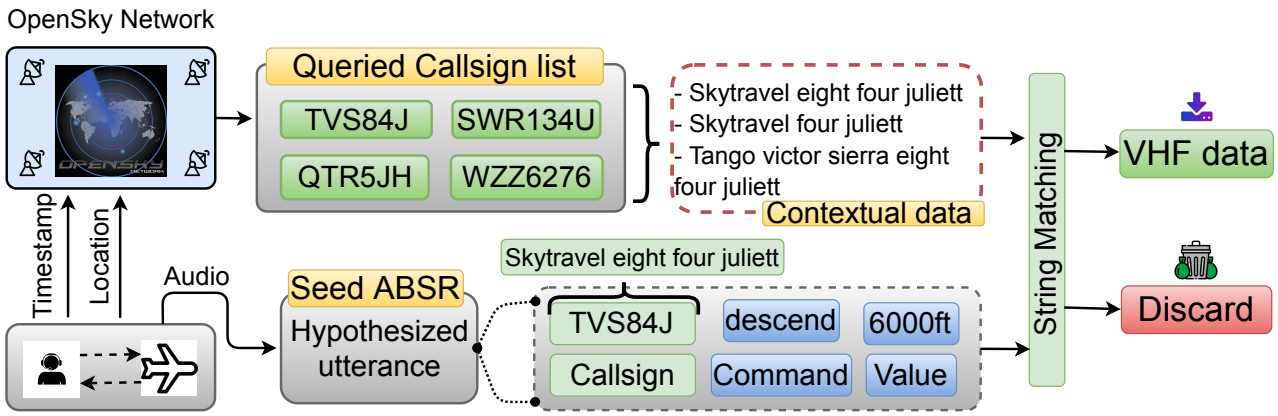

Figure 2. **Pipeline for gathering ATCO-pilot voice communications with VHF receivers**. Timestamp and location of the communication is used to query a callsign list (surveillance data) from OpenSky Network servers. In parallel, the communication is transcribed with an in-domain ABSR system. Later, the contextual data is verbalized, i.e., ICAO callsign → spoken words (red doted box). Finally, a string matching algorithm iteratively tries to match the callsigns in the surveillance data with the words hypothesized by the ABSR, i.e., the transcripts. If a match is found, the segment is stored, otherwise, it is discarded.

not be easy to train. This makes the clustering and the final stage of deciding the speaker role label complicated, and the diarization system might fail. That is one of the main reasons behind our idea to develop a grammar-based speaker role detection based on text (e.g., ASR transcripts).

B. Multi-task ASR

Previous research has shown that to compensate for limited data available in low-resourced languages, multilingual systems are an effective way to train ASR systems [16]–[20]. In such a system, the output layer could be a separate layer for each language, or a single layer shared between all languages [20]. The Kaldi [21] toolkit provides state-of-the-art techniques to train ABSR, specifically, Lattice-Free Maximum Mutual Information (LF-MMI) based models [22]. Recently, [16] showed that multilingual AM can be trained with LF-MMI [22]. In MMI training, the cost function is given as:

$$\mathcal{F}_{\text{MMI}} = \sum_{u=1}^{U} \log \frac{p\left(\mathbf{x}^{(u)}|\mathcal{M}_{\mathbf{w}(u)}, \boldsymbol{\theta}\right) p(\mathbf{w}(u))}{p\left(\mathbf{x}^{(u)}|\mathcal{M}_{\text{den}}, \boldsymbol{\theta}\right)}, \tag{1}$$

where $\mathbf{x}^{(u)}$ is an input sequence for an utterance u, U is a set of all utterances in the training data, $\mathcal{M}_{\mathbf{w}(u)}$ corresponds to a numerator graph specific to a word sequence in transcription, $\mathcal{M}_{\text{den}}$ is a denominator graph modelling all possible sequences which is usually a phone LM, $\boldsymbol{\theta}$ is a model parameter and $p(\mathbf{w}(u))$ is a language model probability for an utterance.

However, in multitask training with separate output layers, the cost function from Equation 1 is computed for each task depending on the number of tasks. For T tasks, the output cost function for each task t depends only on the utterances of that task:

$$\mathcal{F}^{(t)}_{\text{MMI}} = \sum_{u=1}^{U_t} \log \frac{p\left(\mathbf{x}^{(u)}|\mathcal{M}^t_{\mathbf{w}(u)}, \boldsymbol{\theta}\right) p(\mathbf{w}(u))}{p\left(\mathbf{x}^{(u)}|\mathcal{M}^t_{\text{den}}, \boldsymbol{\theta}\right)}, \tag{2}$$

where U_t is the number of utterances in a minibatch for a task t, $\boldsymbol{\theta}$ contains the shared and task-dependent parameters, $\mathcal{M}^t_{\mathbf{w}(u)}$ and $\mathcal{M}^t_{\text{den}}$ are task-specific numerator and denominator graphs, respectively. For a task t, a denominator graph is built using the task-specific phone. For each minibatch, the gradient of each task output layer is computed and updated.

The overall cost-function is then given as a weighted sum of all task-dependent cost-functions defined in Equation 3.

$$\mathcal{F}_{\text{MMI}} = \sum_{t=1}^{T} \alpha_t F^t_{MMI}, \tag{3}$$

where α_t is a task-dependent weight.

Although language and phone sets are the same for ATCO and pilots, due to the variation in the acoustic conditions, we consider them as different tasks and propose to use a multitask approach to train AMs. We hypothesize that using a multitask approach can lead to better ASR performance for both ATCOs and pilots compared to a single AM trained by combining all data.

Different approaches to improve the ASR performance have explored semi-supervised learning [1], [2], [23], integration of surveillance data as prior knowledge into the ASR pipeline [5], [24]–[26] and end-to-end training [27] for ATC domain. Additionally, related work on text-based diarization for ATC communications is explored in [28].

TABLE I. AIR TRAFFIC CONTROL COMMUNICATIONS-RELATED DATABASES USED FOR TRAINING. THE WHOLE DATABASE OF HAAWAII AND ATCO2 WERE NOT AVAILABLE DURING THIS WORK. †TOTAL NUMBER OF AUDIO IN THE DATABASE AFTER SILENCE REMOVAL.

Database	Duration† Training	Open source	Ref
Private databases			
HAAWAII	43	✗	[27]
MALORCA	13	✗	[1], [2]
AIRBUS	100	✗	[29]
Public databases			
ATCOSIM	8	✓	[30]
UWB-ATCC	10.4	✓	[31]
LDC-ATCC	23	✓	[32]
HIWIRE	28.7	✓	[33]
ATCO2	5000	✓	[24]

III. DATASETS

The following subsections provide an overview of the public and private databases used in this paper. A brief overview is also provided in Table I.

A. Collection and Pre-processing of VHF Data

1) Data collection: To obtain ATC voice communications, the following two sources are considered: (i) open-source speech like LiveATC,[7] and ii) speech collected with our own setup of VHF receivers. In addition to speech data, the time-aligned metadata available is used to obtain surveillance data (e.g., callsign list for each communication) from OpenSky Network[8] (OSN). This iterative process yielded ∼377 hours of speech data from Prague (LKPR) and Brno (LKTB) airports from August 2020 until January 2021. This subset is part of a full corpus of around five thousand hours of ATC audio and metadata collected during the ATCO2 project. The full corpus is available for purchase through ELDA in http://catalog.elra.info/en-us/repository/browse/ELRA-S0484. The recordings of both corpus are mono-channel sampled at 16kHz and 16-bit PCM. In this paper, the whole corpus of five thousand hours is not used, as it wasn't available at the time of experimentation.

2) Data pre-processing: Figure 2 shows the pipeline used for preparing the VHF database. First, a seed ASR system is used to produce the transcripts for the 377 hours of collected data. The seed model is a 'hybrid' speech-to-text recognizer based on Kaldi [21] trained with the LF-MMI cost function [22] (see Section II-B). The neural network follows a `cnn-tdnn` topology. It has six `convolutional-layers` followed by nine layers of Factorized Time-Delay Neural Network (`TDNN-F`) [34].

A list of callsigns is retrieved from OSN in ICAO format. Callsigns in ICAO format are composed of three characters airline code, e.g., *TVS*, followed by a flight number which can consist of digits or letters, e.g., *TVS84J*. In order to use this prior knowledge in the ASR, we first verbalize the ICAO callsigns, i.e., transform to the "expanded representation". Several variants exist for a given callsign. As illustrated in Figure 2, the callsign TVS84J can be pronounced as "skytravel eight four juliett" or instead each letter can be spelled out as "tango victor sierra eight four juliett". For more information about these rules, see [24]. Then, an expanded list of callsigns with its variants is created. Finally, string matching of this expanded callsign list is applied to the transcripts generated by the ABSR system. The utterances in which one of the callsigns is found are stored, while the rest are discarded. This pre-processing reduced the data from 377 hours to 66 hours.

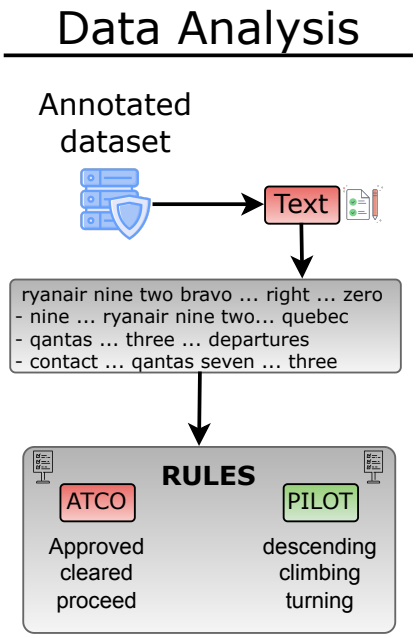

Figure 3. **Overall process to develop a rule-based grammar system to identify ATCOs and pilots based on text**. We perform a data analysis on the already available public and private databases. We then gather a set of words that are more probable to appear in either, ATCO or pilot utterances.

B. Related ATC Datasets Available for Training

In addition to the above data collection, ATCO2 has brought together several air traffic command-related databases [1], [29], [30], [32], [33], [35], [36] from different publicly available open data sources. The full set of databases span approximately 100 hours of speech data that are strongly related in both, phraseology and structure seen in ATCO-pilot

[7]LiveATC.net is a streaming audio network consisting of local receivers tuned to aircraft communications: https://www.liveatc.net/.

[8]OpenSky Network is a non-profit association based in Switzerland. It aims at improving the security, reliability, and efficiency of the airspace usage by providing open access of real-world air traffic control data to the public. The OpenSky Network consists of a multitude of sensors connected to the Internet by volunteers, industrial supporters, and academic/governmental organizations. URL: https://opensky-network.org.

TABLE II. **LIST OF ATCO AND PILOT WORDS USED IN THE PROPOSED GRAMMAR-BASED CLASSIFICATION SYSTEM**. WE COLLECTED 31 WORDS FOR ATCO AND 20 WORDS FOR PILOT.

ATCO words			
approved	back	break	call
cleared	contact	correct	direct
disregard	established	expect	handover
identified	increase	maintain	no
proceed	radar	reduce	report
roger	soon	standby	transition
turn	vortex	wake	wind
you're	you've	yours	
Pilot words			
CPDLC	approaching	climbing	comply
descending	heavy	inbound	maintaining
our	reducing	request	requesting
standing	stopping	taking	turning
us	we	will	wilco

communications [6], [8], [23]. The collection of databases is augmented by adding noise that matched LiveATC audio channels, doubling the size of training data. Additionally, since each of the seven databases had different annotation ontologies (annotation procedure, rules, and symbols), the transcripts had to be standardized and normalized [30], [37].

IV. SPEAKER ROLE CLASSIFICATION WITH TEXT

As described in Section I, to develop a reliable and improved ASR for both, ATCOs and pilots, respective labelled speech data are required. However, in most cases (e.g., ATCO2 project) although large amounts of data are collected, they do not contain speaker labels. The first task is therefore to split the speech recordings into two classes: ATCO and pilot. To accomplish this, we extract the information based on the ICAO grammar to identify the speaker's role. We follow the pipeline described in Figure 3.

ICAO defines a separate grammar for ATCOs and pilots to enable clear communication. For instance, there are certain phrases/commands that an ATCO should use in specific order. This knowledge is used to extract/identify potential words/commands that indicate a specific role of speaker. For example, the words such as "`identified`", "`approved`", "`wind`" would most probably only be spoken by an ATCO and the words "`wilco`", "`maintaining`", "`we`", "`our`" would probably be spoken only by a pilot. Currently, we have made a list of 31 words for ATCO and 20 words for pilot that indicate each role. The list of words is presented in Table II, while the overall pipeline to gather these words is depicted in Figure 3. This list was generated by manual curation and expert feedback. A list of callsigns[9] is also prepared from available airline codes.

[9]The table lists the IATA airline designators, the ICAO airline designators and the airline callsigns (telephony designator). URL: https://en.wikipedia.org/wiki/List_of_airline_codes.

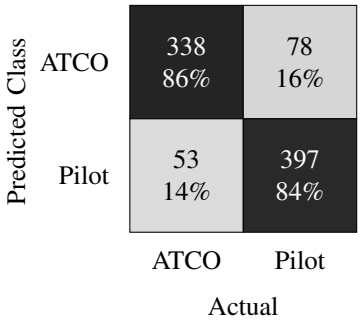

Figure 4. **Confusion matrix for speaker role identification based on text for manually speaker segmented data for London Approach test set**. The total number of ATCO utterances are 391 and the total number of pilot utterances are 475.

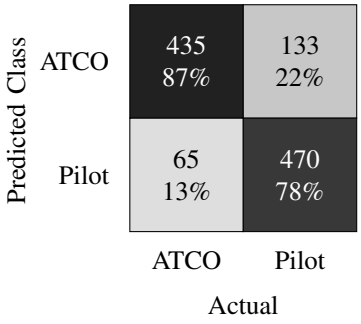

Figure 5. **Confusion matrix for speaker role identification based on text for manually speaker segmented data for Icelandic en-route test set**. The total number of ATCO utterances are 500 and the total number of pilot utterances are 604.

Since this method operates at word level, manual (if available) or automatically generated transcripts are required for the corresponding speech recordings. In order to identify if an utterance is spoken by an ATCO or a pilot, we check the corresponding transcript for the conditions below: if the callsign appears at the beginning of an utterance, this utterance is classified as ATCO, else it is classified as a pilot. As there is a greeting at the beginning quite often, we check if the callsign appears within the first four words. If one of the words in the utterance is in the list of ATCO words or in the list of pilot words, then the respective role is assigned.

Once each utterance in the training data is tagged with ATCO or pilot labels, we propose to train two versions of ASR. In the first system, there are two acoustic models: one for ATCO and one for pilot. In the second system, we train a multitask network with one task as ATCO ASR and the other as pilot ASR (see Section II-B). The procedure is illustrated in Figure 1.

A. Assigning Scores to Decisions

The grammar role also provides the probability of assigning a speaker role to a given utterance using the bag-of-words that are manually created. In order to obtain such probability,

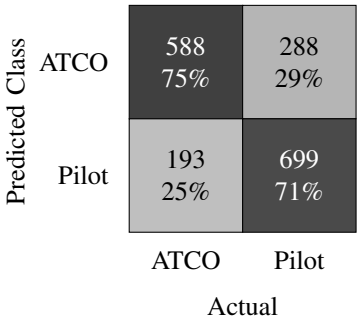

Figure 6. **Confusion matrix for speaker role identification based on text for manually speaker segmented data for LiveATC data**. The total number of ATCO utterances are 781 and the total number of pilot utterances are 987.

Bayes' rule is adopted. For instance, the probability of an utterance being ATCO is computed as:

$$p(\text{atco}|\text{utt}) = \frac{p(\text{utt}|\text{atco})p(\text{atco})}{p(\text{utt}|\text{atco})p(\text{atco}) + p(\text{utt}|\text{pilot})p(\text{pilot})}, \quad (4)$$

Here $p(\text{atco})$ and $p(\text{pilot})$ are the priors, and we assume both classes have equal probability and hence their value is 0.5. The $p(\text{utt}|\text{atco})$ is computed as:

$$p(\text{utt}|\text{atco}) = \prod_{w_i \in \text{utt}} p(w_i|\text{atco}). \quad (5)$$

Similarly, the $p(\text{utt}|\text{pilot})$ is computed as:

$$p(\text{utt}|\text{pilot}) = \prod_{w_i \in \text{utt}} p(w_i|\text{pilot}). \quad (6)$$

The $p(w_i|\text{atco})$ and $p(w_i|\text{pilot})$ are computed from using the 15k speaker role annotated utterances available as part of HAAWAII project from the Air Navigation Service Providers (ANSPs) for training: i) NATS for London Approach and ii) ISAVIA for Icelandic en-route where the total number of utterances for ATCO and pilot are 7k and 8k respectively. The below equation is used to compute this:

$$p(w_i|\text{class}) = \frac{\text{class count}}{\text{total count}}, \quad (7)$$

where class count is the number of times the word w_i appears in that particular class, and total count is the sum of the number of times the words in both the classes.

TABLE III. **COMPARISON OF WORD ERROR RATES (WER) IN PERCENTAGES FOR ACOUSTIC MODELS TRAINED WITH DATA FROM OTHER ATC DATASETS**. THE MODELS ARE TESTED ON LIVEATC ATCO AND PILOT TEST SETS. THE RESULTS SHOW THAT TRAINING SPEAKER-DEPENDENT ACOUSTIC MODELS OR A MULTITASK SYSTEM PROVIDE BETTER ASR PERFORMANCE THAN THE COMBINED SYSTEM.

Model	**WER %**	
	ATCO test	**Pilot test**
Clean	36.9	47.7
Noise	**31.3**	**41.1**
Combined	36.1	45.8
Multitask	31.6	**41.1**

B. Speaker Role Classification Performance

This method has been tested on manually speaker segmented and transcribed data for three different test sets: i) NATS for London Approach, ii) ISAVIA for Icelandic en-route and iii) LiveATC test set. In the first set, there are 391 and 475 ATCO and pilot utterances, respectively. From the confusion matrix shown in Figure 4, we can observe that this method provides a true positive rate (TPR) of 86% (correctly classified ATCO) and true negative rate (TNR) of 84% (correctly classified pilot). The second set used consists of 500 ATCO utterances and 604 pilot utterances. From the confusion matrix shown in Figure 5, we see that this method provides a TPR of 87% and TNR of 78%. For the third set, we see a TPR of 75% and a TNR of 71%. This shows that the bag-of-words generated match the first two sets and the communication is slightly different since there is a domain mismatch caused by data from different airports.

C. Error Analysis

As there exists many variants for any given callsign, checking only for the airline code (e.g., lufthansa) is a major factor contributing to the misclassification of ATCO as pilot. A reason for the misclassification of pilot as ATCO is the occurrence of callsigns at the beginning of the utterance. Analysis of misclassification errors show that the accuracy can be improved by i) matching the callsign spoken with its allowed variants[10] and ii) using the context prior to the callsigns.[11] We will consider applying the aforementioned improvements as a part of our future work.

V. EXPERIMENTS

For all our experiments, conventional biphone Convolutional Neural Network (CNN) [38] + TDNN-F [34] based acoustic

[10]For instance, LUF189AF → lufthansa one eight nine alfa foxtrot, one eight nine alfa foxtrot, etc.

[11]An example is that the pilot may mention the place of the control they want to communicate followed by the callsign

TABLE IV. COMPARISON OF WORD ERROR RATES (WER) IN PERCENTAGE FOR ACOUSTIC MODELS TRAINED WITH ONLY THE DATA COLLECTED FROM VHF RECEIVERS. THE MODELS ARE TESTED ON LIVEATC ATCO AND PILOT TEST SETS.

Model	WER %	
	ATCO test	Pilot test
VHF ATCO	43.2	51.6
VHF Pilot	40.3	45
Combined	46	50
Multitask	**38.2**	**44**

TABLE V. COMPARISON OF WORD ERROR RATES (WER) IN PERCENTAGES FOR ACOUSTIC MODELS TRAINED WITH ALL ATCO AND PILOT DATA FROM ALL DATABASES. ADDITIONALLY, THE TRAINING DATA IS AUGMENTED WITH NOISE.

Model	WER %	
	ATCO test	Pilot test
ATCO	**30.3**	43.2
Pilot	32.8	**40.3**
Combined	31.2	41.3
Multitask	31.9	41.3

models trained with Kaldi [21] toolkit (i.e., nnet3 model architecture) is used. AMs are trained with the LF-MMI [22] training framework, considered to produce state-of-the-art performance for hybrid ASR systems. In all the experiments, 3-fold speed perturbation [39] and i-vectors features are used. The multitask training script used can be found in Kaldi [21].[12] The value of the task dependent weight α_t used in our experiments is 0.5. Language model (LM) is trained with all the manual transcripts available from datasets described in Section III-B and used for all the experiments.

The performance of different models is evaluated on LiveATC test set with the Word Error Rate (WER) metric. WER is computed with the Levenshtein distance at the word level. The total duration of the test set is 1h 50 mins. The set is split into two subsets: ATCO set (52 mins) and Pilot set (58 mins). In each group of experiments, results are given for i) AM trained for each task separately, ii) AM trained by combining all data and iii) AM trained with multitask learning.

A. Experiments on ATC Databases

In this setup, we use data from the ATC databases mentioned in Section III-B as Clean data and its noise augmented part as Noise data. In this setup, the data is not split to ATCO and pilot. As shown in Table III, both ATCO and pilot test sets provide better performance when the model is trained with Noise data compared to the model trained with only Clean data. This shows that the noise augmented version of the clean data matches with the test sets much better than the clean version. Moreover, the Combined system performs significantly worse than the Noise system. This shows that using the Clean dataset in fact hurts ASR performance. This is one of the reasons why the multitask system performs only on par with the Noise system. Thus, only the noise augmented data is used for training in the next experiments.

[12]The script in located in: https://github.com/kaldi-asr/kaldi/blob/master/egs/babel/s5d/local/chain2/run_tdnn.sh.

B. Experiments on VHF Data

Results in Table IV are presented for AMs trained with only the VHF data. Applying speaker role identification for the pre-processed data (66 h) yields 43 h for ATCO and 23 h for Pilot. Similar to Table III, the results in Table IV show that using multitask learning instead of training AM by combining all the data provides better ASR performance. Furthermore, the results reveal that due to the low amount of data, multitask learning outperforms its single task counterparts.

C. Experiments on VHF and Other ATC Datasets

In this subsection, we report results with models trained from both VHF and ATC datasets used in the previous two experiments. By investigating the ATC databases used in Section V-A, we discovered that some datasets also contain pilot speech. Since no speaker role labels are available for these sets, we applied the proposed method to split the noise augmented speech as ATCO or pilot and combined them with their respective classes of the VHF data. This provided 123h of data for ATCO and 80h for pilot. The results in Table V show that training AMs for each task separately performs relatively better, by 2.9% for ATCO and 2.4% for pilot, than using the Combined system. This suggests that when more data is available, using our grammar-based approach to obtain speaker role information to train separate ATCO and pilot ASR is better than the Combined approach. The multitask system does not perform better than the Combined one. This means that there is a negative transfer when considering ATCO and pilot tasks. This is expected as the ATC data dominates in size during training.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we compared different types of training AMs with state-of-the-art LF-MMI framework for ATCO and pilot speech recordings. The developed ASR systems were evaluated separately on ATCO and pilot test sets built from LiveATC. Due to the noisy nature of both ATCO and pilot test sets, AM trained with only noise augmented speech data boosts the ASR performance. We proposed a simple grammar

based approach to identify speaker roles automatically and train acoustic models either by speaker role or in a multitask fashion. The results show that multitask training approach outperforms other training methods when limited training data is available. When sufficient data is available, we show that training AMs separately provides better ASR performance for both ATCO and pilot compared to the model trained by combining all data. Relative improvements of 3.2% for the ATCO set and 1.9% for the pilot set were obtained.

As mentioned earlier, the rule-based approach can further be improved by taking into account all the allowed variants of a callsign and using the context prior to the callsigns during classification. In our current work, we explored only the acoustic modeling part of speech recognizer. As a part of our future work, we consider investigating the improvement of speaker-dependent ASR systems by i) training a separate LM for each speaker class or ii) interpolating the class specific LM with the baseline LM.

ACKNOWLEDGEMENTS

The work was supported by SESAR EC project No. 884287 - HAAWAII (Highly automated air-traffic controller workstations with artificial intelligence integration). The work was also partially supported by the European Union's Horizon 2020 project No. 864702 - ATCO2 (Automatic collection and processing of voice data from air-traffic communications), which is a part of Clean Sky Joint Undertaking. We wish to acknowledge Santosh Kesiraju for providing valuable insights and suggestions regarding the assignment of scores for classification.

REFERENCES

[1] A. Srinivasamurthy, P. Motlícek, I. Himawan *et al.*, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2406–2410. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1446.html

[2] M. Kleinert, H. Helmke, G. Siol *et al.*, "Semi-supervised adaptation of assistant based speech recognition models for different approach areas," in *37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018.

[3] 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.

[4] 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.

[5] M. Kocour, K. Veselý, A. Blatt *et al.*, "Boosting of contextual information in ASR for air-traffic call-sign recognition," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget *et al.*, Eds. ISCA, 2021, pp. 3301–3305. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1619

[6] J. Zuluaga-Gomez, K. Veselỳ, A. Blatt *et al.*, "Automatic call sign detection: Matching air surveillance data with air traffic spoken communications," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 59, no. 1, 2020, p. 14.

[7] M. Rigault, C. Cevenini, K. Choukri *et al.*, "Legal and ethical challenges in recording air traffic control speech," in *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 79–83. [Online]. Available: https://aclanthology.org/2022.legal-1.14

[8] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan *et al.*, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," in *Interspeech*, 2020, pp. 2297–2301.

[9] 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 *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2993–2997. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1962

[10] S. Ruder, "An overview of multi-task learning in deep neural networks," *ArXiv preprint*, vol. abs/1706.05098, 2017. [Online]. Available: https://arxiv.org/abs/1706.05098

[11] X. Anguera, S. Bozonnet, N. Evans *et al.*, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[12] T. J. Park, N. Kanda, D. Dimitriadis *et al.*, "A review of speaker diarization: Recent advances with deep learning," *ArXiv preprint*, vol. abs/2101.09624, 2021. [Online]. Available: https://arxiv.org/abs/2101.09624

[13] J. Zuluaga-Gomez, K. Veselỳ, I. Szöke *et al.*, "ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications," *arXiv preprint arXiv:2211.04054*, 2022.

[14] ALLCLEAR, "Icao phraseology reference guide," 2020. [Online]. Available: https://www.skybrary.aero/bookshelf/books/115.pdf

[15] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101254, 2022.

[16] S. R. Madikeri, B. K. Khonglah, S. Tong *et al.*, "Lattice-free maximum mutual information training of multilingual speech recognition systems," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4746–4750. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2919

[17] L. Burget, P. Schwarz, M. Agarwal *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4334–4337.

[18] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech communication*, vol. 56, pp. 142–151, 2014.

[19] N. T. Vu, D. Imseng, D. Povey *et al.*, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 7639–7643.

[20] M. Karafiát, M. K. Baskar, P. Matějka *et al.*, "Multilingual blstm and speaker-specific vector adaptation in 2016 but babel system," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 637–643.

[21] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[22] D. Povey, V. Peddinti, D. Galvez *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 2751–2755. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-595

[23] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad *et al.*, "Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and

Untranscribed ATC Data in ASR Systems," in *Interspeech*, 2021, pp. 3296–3300.

[24] M. Kocour, K. Veselỳ, I. Szöke *et al.*, "Automatic processing pipeline for collecting and annotating air-traffic voice communication data," *Engineering Proceedings*, vol. 13, no. 1, p. 8, 2021.

[25] I. Nigmatulina, R. Braun, J. Zuluaga-Gomez, and P. Motlicek, "Improving callsign recognition with air-surveillance data in air-traffic communication," *ArXiv preprint*, vol. abs/2108.12156, 2021. [Online]. Available: https://arxiv.org/abs/2108.12156

[26] I. Nigmatulina, J. Zuluaga-Gomez, A. Prasad *et al.*, "A two-step approach to leverage contextual data: speech recognition in air-traffic communications," in *ICASSP*, 2022.

[27] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina *et al.*, "How does pre-trained wav2vec2.0 perform on domain shifted asr? an extensive benchmark on air traffic control communications," *IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar*, 2023.

[28] J. Zuluaga-Gomez, S. S. Sarfjoo, A. Prasad *et al.*, "Bertraffic: Bert-based joint speaker role and speaker change detection for air traffic control communications," *IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar*, 2023.

[29] E. Delpech, M. Laignelet, C. Pimm *et al.*, "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)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. [Online]. Available: https://aclanthology.org/L18-1453

[30] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/545_paper.pdf

[31] L. Šmídl, J. Švec, D. Tihelka *et al.*, "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.

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

[33] J. Segura, T. Ehrette, A. Potamianos *et al.*, "The hiwire database, a noisy and non-native english speech corpus for cockpit communication," *Online. http://www. hiwire. org*, 2007.

[34] D. Povey, G. Cheng, Y. Wang *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3743–3747. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1417

[35] L. Šmídl, J. Švec, D. Tihelka *et al.*, "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.

[36] S. Pigeon, W. Shen, A. Lawson, and D. A. v. Leeuwen, "Design and characterization of the non-native military air traffic communications database (nnmatc)," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[37] H. Helmke, M. Slotty, M. Poiger *et al.*, "Ontology for transcription of atc speech commands of SESAR 2020 solution PJ.16-04," in *IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018.

[38] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[39] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.