GRAMMAR BASED IDENTIFICATION OF SPEAKER ROLE FOR IMPROVING ATCO AND PILOT ASR

Amrutha Prasad        Zuluaga-Gomez, Juan
Petr Motlicek          Oliver Ohneiser        Hartmut Helmke
Seyyed Saeed Sarfjoo   Nigmatulina Iuliia

Idiap-RR-22-2021

Version of DECEMBER 20, 2021
Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition

Amrutha Prasad, Juan Zuluaga-Gomez, Petr Motlicek, Saeed Sarfjoo, Iuliia Nigmatulina
Idiap Research Institute
Martigny, Switzerland
{amrutha.prasad,juan-pablo.zuluaga,petr.motlicek, saeed.sarfjoo,iuliia.nigmatulina}@idiap.ch

Oliver Ohneiser, Hartmut Helmke, German Aerospace Center (DLR), Institute of Flight Guidance
Braunschweig, Germany
{oliver.ohneiser,hartmut.helmke}@dlr.de

Abstract—Assistant Based Speech Recognition (ABSR) for air traffic control is generally trained by pooling both Air Traffic Controller (ATCO) and pilot data. In practice, this is motivated by the fact that the proportion of pilot data is lesser compared to ATCO while their standard language of communication is similar. However, due to data imbalance of ATCO and pilot and their varying acoustic conditions, the ASR performance is usually significantly better for ATCOs than pilots. In this paper, we propose to (1) split the ATCO and pilot data using an automatic approach exploiting ASR transcripts, and (2) consider ATCO and pilot ASR as two separate tasks for Acoustic Model (AM) training. For speaker role classification of ATCO and pilot data, a hypothesized ASR transcript is generated with a seed model, subsequently used to classify the speaker role based on the knowledge extracted from grammar defined by International Civil Aviation Organization (ICAO). This approach provides an average speaker role identification accuracy of 83% for ATCO and pilot. Finally, we show that training AMs separately for each task, or using a multitask approach is well suited for this data compared to AM trained by pooling all data.

Keywords—assistant based speech recognition, air traffic management, multitask acoustic model, speaker classification

I. INTRODUCTION

Previous research [1], [2] as part of the MALORCA\textsuperscript{1} and AcListent-Strips\textsuperscript{2} project, respectively, focused on i) improving ABSR accuracy for ATCOs, ii) reducing workload for ATCOs [3], and iii) increasing efficiency [4] of ATCOs. As part of an ongoing HAAWAI\textsuperscript{3} project, we aim to research and develop a reliable and adaptable solution to automatically transcribe voice commands issued by both ATCOs and pilots.

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 another ongoing project called ATCO2\textsuperscript{4} is to collect large set of voice recordings of ATCOs and pilots (with a minimum effort) for the aforementioned purpose. In order to train ASR for this task, ATCO and pilot speech recordings are usually pooled together [1], [5], [6] 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 performance is significantly different if applied on ATCO or pilot speech (i.e. ATCO’s speech is easier to recognize). Our baseline system trained by pooling all data reveals that the absolute difference in WER for ATCO and pilot is 9.7% (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 [7].

The classification of speaker roles is not only important to improve ASR quality. It also improves succeeding natural language processing tasks, i.e., it enhances automatic annotation of extracted ATC commands from transcripts. A developed and European-wide agreed ontology [8] distinguishes between ATCO and pilot utterances to accurately recognize different elements of ATC commands in a tower environment [9] or for read back error detection in en-route environment [10].

In this paper, we hypothesize that instead of developing the ASR as a single task, ATCO and pilot ASR can be considered as two separate tasks [11]. Specifically, this paper investigates a multitask approach to train AMs to be integrated in ASR for ATCO and pilot. An obvious first step is to automatically split the ATC speech communications into two tasks (i.e. obtaining these speaker labels manually on a large dataset would be expensive and time consuming). A common approach is to use speaker diarization to classify the speakers in the audio [12], [13]. Although the ATCO speech is often cleaner than the pilot (as the former communicates from a controlled acoustic environment), the speech recordings collected in ATCO2

\textsuperscript{1}MACHINE Learning Of speech Recognition models for Controller Assistance: http://www.malorca-project.de/wp/
\textsuperscript{2}Active Listening Assistant Strips: https://www.malorca-project.de/wp/
\textsuperscript{3}Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration: https://www.haaawi.de
\textsuperscript{4}Automatic collection and processing of voice data from air-traffic communications https://www.atco2.org/
project using Very High Frequency (VHF) receivers are noisy for both ATCO and pilot channels. In such a case, the speaker diarization system may fail to assign speaker labels (ATCO or pilot) accurately. Thus, a speaker diarization system cannot be easily deployed to obtain accurate speaker labels.

The vital aspect in the air traffic management (ATM) environment is the communication between a controller and pilot. For the smooth travel of the aircraft this communication is well defined with a standard phraseology by ICAO [14]. Another approach to obtain the speaker class is through leveraging the ‘ICAO’ grammar to classify an utterance as one of the classes on the text level. Once the speaker labels (ATCO and pilot) are available for the large data, AMs can be trained for both controllers and pilots through different approaches. In this study, we show that due to the poor acoustic conditions training a single AM by pooling all data does not provide the best performance for pilots even if the speech is constrained by grammar. To obtain better performance accuracy, AM should be trained separately for ATCO and pilot data or considered as different tasks by using a multitask approach.

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

Previous research [15]–[19] has shown that to compensate for limited data available in low-resourced languages, multilingual systems are an effective way to train ASR systems. In such a system, the output layer could be a separate layer for each language, or a single layer shared between all languages [19]. The Kaldi [20] toolkit provides state-of-the-art techniques to train AMs, specifically Lattice-Free Maximum Mutual Information (LF-MMI) [21]. Recently, [15] showed that multilingual AM can be trained with LF-MMI [21]. In MMI training, the cost function is given as:

$$F_{MMI}^{(t)} = \sum_{u=1}^{U_t} \frac{\log p\left(x^{(u)}|M_w^{(u)}, \theta\right) p(w(u))}{p\left(x^{(u)}|M_{\text{den}}, \theta\right)},$$

where $U_t$ is the number of utterances in a minibatch for a task $t$, $\theta$ contains the shared and task-dependent parameters, $M_w^{(u)}$ and $M_{\text{den}}$ 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.

$$F_{MMI} = \sum_{t=1}^{T} \alpha_t F_{MMI}^{(t)},$$

where $\alpha_t$ 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.

III. DATASETS

The following subsections provide an overview of the data used in this paper.

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\(^3\), 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 the contextual information (e.g. callsign list for each utterance) from the OpenSky Network\(^6\) (OSN). This process yielded 377 hours of speech data from Prague (LKPR) and Brno (LKTB) airports from August 2020 until January 2021 for ATCO2 project.

\(^3\)LiveATC.net is a streaming audio network consisting of local receivers tuned to aircraft communications: https://www.liveatc.net/

\(^6\)OpenSky Network: provides open access of real-world air traffic control data to the public
2) Data pre-processing: Figure 1 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 [20] trained with the LF-MMI cost function [21]. The neural network has six convolutional layers followed by nine Factorized Time-Delay Neural Network (TDNN-F) [22].

A list of callsigns is retrieved from OSN in ICAO format. The ICAO format for a callsign is composed of three characters airline code (e.g., TVS) followed by a flight number which can consist of digits or letters, e.g. leading to TVS84J. In order to use this prior knowledge, this format is transformed into its "expanded version". Several variants exist for a given callsign. As illustrated in Figure 1, the callsign TVS84J can be pronounced as "skytravel eight four juliett" or instead each letter can be spelt out "tango victor sierra eight four juliett".

Then, an ensemble of callsigns with its variants are created. Finally, string matching of this expanded callsign list is applied to the automatic transcripts. The utterances in which one of the callsigns is found are stored. This pre-processing reduced the data from 377 hours to 66 hours.

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], [23]–[27] from different publicly available open data sources. The full set of databases span approximately 140 hours of speech data that are strongly related in both phraseology and structure seen in ATCO-pilot communications [5], [6], [28]. These databases were additionally augmented by adding noises that match LiveATC audio channels, doubling the size of training data. Since each of the seven databases had different annotation ontologies (annotation procedure, rules, and symbols), the transcripts had to be standardized and normalized [8], [25].

IV. Speaker Role Classification with Text

As described in Section 1, to develop a reliable and better performing ASR for both ATCOs and pilots, respective labelled speech data are required. However, in most cases, e.g., such as in 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.

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 a 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 21 words for pilot that indicate each role. The list of words are presented in Table I. This list was generated by manual curation and expert feedback. A list of
callsigns\textsuperscript{7} is also prepared from available airline codes.

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 classified as ATCO or pilot, 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 other as pilot ASR. The procedure is illustrated in Figure 2.

\textsuperscript{7}\url{https://en.wikipedia.org/wiki/List_of_airline_codes}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|}
\hline
\textbf{ATCO words} & \textbf{Pilot words} & \multicolumn{2}{c|}{\textbf{Predicted Class}} \\
\hline
approved & back & ATCO & 435 \\
clared & contact & Pilot & 65 \\
disregard & correct & ATCO & 588 \\
identified & direct & Pilot & 193 \\
proceed & radar & ATCO & 470 \\
roger & soon & Pilot & 699 \\
turn & standby & ATCO & 435 \\
you're & wake & Pilot & 133 \\
yours & wind & ATCO & 53 \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Confusion matrix for speaker role identification based on text for manually speaker segmented data for London Approach test set. Total number of ATCO utterances are 391 and the total number of pilot utterances are 475.}
\end{figure}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
\textbf{Predicted Class} & \textbf{Actual} & \\
\hline
ATCO & 338 & 78 \\
    & 86\% & 16\% \\
Pilot & 53  & 397 \\
    & 14\% & 84\% \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Confusion matrix for speaker role identification based on text for manually speaker segmented data for Icelandic en-route test set. Total number of ATCO utterances are 500 and the total number of pilot utterances are 604.}
\end{figure}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
\textbf{Predicted Class} & \textbf{Actual} & \\
\hline
ATCO & 588 & 288 \\
    & 75\% & 29\% \\
Pilot & 193 & 699 \\
    & 25\% & 71\% \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Confusion matrix for speaker role identification based on text for manually speaker segmented data for LiveATC data. Total number of ATCO utterances are 781 and the total number of pilot utterances are 987.}
\end{figure}

\subsection{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 Bayes’ rule is adopted. For e.g., the probability of an utterance being ATCO is computed as:

\begin{equation}
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})}
\end{equation}

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

\begin{equation}
p(\text{utt}|\text{atco}) = \prod_{w_i \in \text{utt}} p(w_i|\text{atco}).
\end{equation}

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

\begin{equation}
p(\text{utt}|\text{pilot}) = \prod_{w_i \in \text{utt}} p(w_i|\text{pilot})
\end{equation}
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 HAAWAI project from the Air Navigation Service Providers (ANSPs) for training: i) NATS for London Approach and ii) ISA VIA 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 number of times the words in both the classes.

**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) ISA VIA for Icelandic en-route and iii) LiveATC test set. In the first set, there are 391 ATCO utterances and 475 pilot utterances. From the confusion matrix shown in Figure 3, 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 4, 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 are different airports and the communication is different.

**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 (e.g. LUF189AF → lufthansa one eight nine alfa foxtrot, one eight nine alfa foxtrot, etc) and ii) using the context prior to the callsigns (e.g., the pilot may mention the place of the control they want to communicate followed by the callsign). 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) [29] + TDNN-F [22] based acoustic models trained with Kaldi [20] toolkit (i.e. nnet3 model architecture) is used. AMs are trained with the LF-MMI [21] training framework considered to produce state-of-the-art performance for hybrid ASR systems. In all the experiments, 3-fold speed perturbation [30] and i-vectors are used. The multi-task training script used can be found in Kaldi [20]8. The value of the task dependent weight $\alpha_t$ 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 which is based on 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. As shown in Table II, 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

---

8egs/babel multilang/s5d/local/chain2/run_tdnn.sh
TABLE IV. WER COMPARISON FOR MODELS TRAINED WITH ALL ATCO DATA FROM ALL DATABASES AND ALL PILOT DATA WITH NOISE AUGMENTED DATA.

<table>
<thead>
<tr>
<th>Model</th>
<th>WER %</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ATCO</td>
<td>Pilot</td>
<td></td>
</tr>
<tr>
<td>ATCO</td>
<td>30.3</td>
<td>43.2</td>
<td></td>
</tr>
<tr>
<td>Pilot</td>
<td>32.8</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>31.2</td>
<td>41.3</td>
<td></td>
</tr>
<tr>
<td>Multitask</td>
<td>31.9</td>
<td>41.3</td>
<td></td>
</tr>
</tbody>
</table>

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 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 separate LM for each speaker class or ii) interpolating the class specific LM with the clean data.

ACKNOWLEDGEMENTS

The work was supported by SESAR EC project No. 884287 - HAAWAI (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


