Automatic Speech Analysis Framework for ATC Communication in HAAWAII

Petr Motlicek[1,2†], Amrutha Prasad[1,2†], Iuliia Nigmatulina[1,3],
Hartmut Helmke[4], Oliver Ohneiser[4], Matthias Kleinert[4]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Brno University of Technology, Speech@FIT, Brno, Czech Republic
[3]Institute of Computational Linguistics, University of Zurich, Switzerland
[4]Institute of Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany
[†]equal contribution

***Abstract*—Over the past years, several SESAR funded exploratory projects focused on bringing speech and language technologies to the Air Traffic Management (ATM) domain and demonstrating their added value through successful applications. Recently ended HAAWAII project developed a generic architecture and framework, which was validated through several tasks such as callsign highlighting, pre-filling radar labels, and readback error detection. The primary goal was to support pilot and air traffic controller communication by deploying Automatic Speech Recognition (ASR) engines. Contextual information (if available) extracted from surveillance data, flight plan data, or previous communication can be exploited via entity boosting to further improve the recognition performance. HAAWAII proposed various design attributes to integrate the ASR engine into the ATM framework, often depending on concrete technical specifics of target air navigation service providers (ANSPs). This paper gives a brief overview and provides an objective assessment of speech processing components developed and integrated into the HAAWAII framework. Specifically, the following tasks are evaluated w.r.t. application domain: (i) speech activity detection, (ii) speaker segmentation and speaker role classification, as well as (iii) ASR. To our best knowledge, HAAWAII framework offers the best performing speech technologies for ATM, reaching high recognition accuracy (i.e., error-correction done by exploiting additional contextual data), robustness (i.e., models developed using large training corpora) and support for rapid domain transfer (i.e., to new ATM sector with minimum investment). Two scenarios provided by ANSPs were used for testing, achieving callsign detection accuracy of about 96% and 95% for NATS and ISAVIA, respectively.**

***Keywords*—HAAWAII project, Speech activity detection, Speaker segmentation, Speaker role classification, Automatic Speech Recognition.**

I. Introduction

During the last decade, many successful applications of Automatic Speech Recognition and Understanding (ASRU) for Air Traffic Management (ATM) have been developed. Supporting Air Traffic Controllers (ATCos) by pre-filling radar label entries with ASRU has achieved a Technology Readiness Level (TRL) of 6, which was validated in SESAR 2020 funded industrial research [1]. Yet, the air-traffic control (ATC) communication remains technologically challenging domain due to its high variability (different airports, accents, noise, etc.), lack of data, and high level of responsibility. Moreover, many ASRU applications require a real-time processing capability. Although offline processing of ATC communication can find many applications (e.g., workload prediction does not require fast response and thus offline or even optimized batch processing can be well integrated), the real-time processing (i.e., a response is typically a few tens or hundreds of milliseconds) cannot be avoided in the case of pre-filling radar labels or readback error detection applications.

As different ASRU applications developed for ATC can have different requirements, different architectures and modules need to be addressed from the early development stages. Previous research on Automatic Speech Recognition (ASR) for ATC [2], [3], mainly focused on reduction of ATCo's workload [4], improvements of ATM efficiency [5] or machine learning of models for adaptation to new ATM environments as in the MALORCA[1] project. The need for a flexible and technologically advanced framework for the ATC domain in general became the main motivation for the HAAWAII (**H**ighly **A**dvanced **A**ir traffic controller **W**orking position With **A**rtificial **I**ntelligence **I**ntegration) project[2]. More precisely, following crucial questions were addressed by HAAWAII:

- How to automatically and reliably detect in near real-time the start and the end of communication, especially in cases when the technological solutions deployed by air navigation service providers (ANSPs) do not enable to extract Push-To-Talk (PTT) signal?
- How to develop and integrate advanced real-time ASR engines offering sufficient recognition accuracies acceptable for their ATM deployment?
- How to automatically detect who speaks, specifically in cases where inbound and outbound ATC communication (i.e., from ATCo or from pilot) channels are combined to one due to the hardware constraint deployed at ANSP side?

Supported by SESAR Joint Undertaking, grant agreement No 884287

[1]https://www.malorca-project.de/
[2]https://www.haawaii.de/wp/

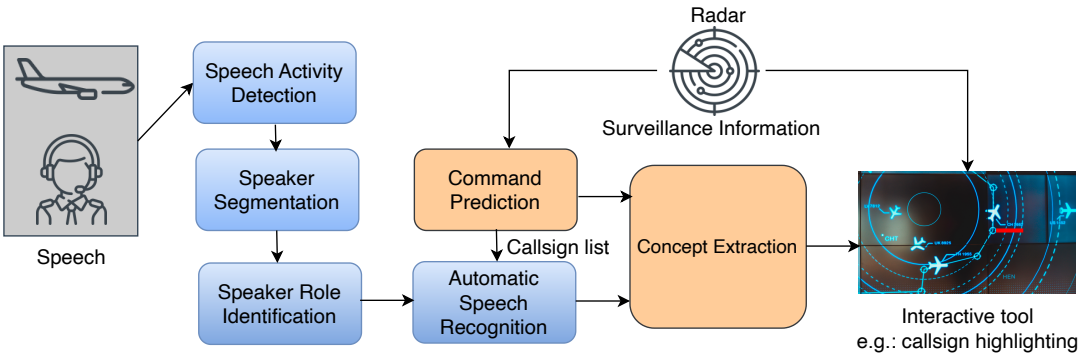

Figure 1. HAAWAII framework.

- How do the ASR errors propagate to the downstream tasks such as Text-to-Concept extraction (recent work for HAAWAII described in [6])?
- How to accurately detect callsigns, as its importance is above other mentioned tasks?
- How can the complementary information (e.g., contextual data available from surveillance data) be exploited to reduce the search space through ASR decoders?

The rest of the paper is organized as follows. Section II describes the individual processing blocks of the HAAWAII framework. The objective assessment of these blocks is presented in Section III. This is followed by Conclusions in Section IV.

II. HAAWAII Framework

The HAAWAII framework is a system designed to transform ATC speech communication into conceptual elements that can be seamlessly integrated into various ATM applications. This framework takes into account both the transmissions from ATCos and pilots.

As seen in Figure 1, the HAAWAII framework is composed of two sets of blocks, namely related to speech recognition (blue) and understanding (orange). Figure 1 illustrates different modalities, i.e., speech communication and surveillance, are utilized across various blocks. This paper primarily addresses speech technologies developed in HAAWAII, while the comprehensive overview of speech understanding block (i.e., semantics) is given in [6]. These speech processing blocks encompass Speech Activity Detection (SAD), Speaker Segmentation, Speaker Role Detection, ASR and its enhancement by applying contextual information (boosting), and ASR adaptation.

A. Speech Activity Detection

Speech Activity Detection (SAD) serves as a critical component integrated to the conventional speech-processing systems, Goal of SAD is to reliably detect the beginning and end of the speech in the input signal. In case of real-time processing, the detection can trigger subsequent processing blocks such as ASR. In case of offline processing, detected boundaries can be used to segment input stream into a sequence of ”speech utterances” (in case of ATC communication typically a few seconds long). Good quality SAD is indispensable for both offline and online streaming recognition applications.

More specifically, in case of offline processing, SAD is used to segment large amounts of ATC communication data (typically provided by ANSPs) further used for training and evaluating the developed technologies. Conversely, in real-time online scenarios, SAD triggers subsequent processing tasks if the speech is detected in the communication between ATCos and pilots. The task becomes more challenging in case of detecting the pilot speech which is often obscured by background noise. High-quality SAD becomes especially crucial when the Push-To-Talk (PTT) signal is not available due to technical constraints.

a) Evaluated SAD technologies: Various SAD technologies were explored and tested as part of HAAWAII project to reliably segment ATC speech. The inspiration came from the work by S. Sarfjoo et al. [7]. Namely, following models were tested for ATC: (1) BUT’s phoneme recognizer-based SAD (so called Phn Rec), (2) Google WebRTC, (3) Generic neural model trained using multi-lingual ASR, (4) Bi-directional LSTM (BLSTM) neural model developed by Pyannote, and (5) Kaldi energy-based SAD:

- ”Phn Rec” was developed on top of a Hungarian phoneme recognizer where all the non-silent classes are linked to the final ”speech” class [8]. Hungarian speech data which was collected in SpeechDat-E project[3], was found as the best for generic phoneme recognition working across many languages [9].
- Google WebRTC is an open framework for the web that enables real-time communications capabilities in the browser. SAD module can be used separately for the speech/non-speech detection[4].
- SAD built on top of neural model trained from multi-lingual speech: To investigate the generalization abilities of SAD across noisy speech and for more languages, a multi-lingual solution was developed. Specifically, a multi-lingual acoustic model [10] was trained with lattice-

[3]http://www.fee.vutbr.cz/SPEECHDAT-E
[4]https://webrtc.org

free maximum mutual information criterion [11] using 18 BABEL languages[5]. The pseudo log-likelihoods obtained from this model are used for speech/non-speech detection. To obtain a single decision, the following fusion approaches were considered:
  - The best score across all languages is used (ASR SingleBest).
  - The outputs from all languages are fused using logistic regression (ASR Mul LR).
  - The outputs are fused using a majority voting approach (ASR Mul MV).
- The production pyannote SAD exploiting LSTM based neural architecture was trained by using the same 18 BABEL languages.
- The energy based SAD exploits energy extracted from each individual frame of the input signal to make a speech/non-speech decision.

B. Speaker Segmentation

Speaker segmentation, also known as speaker diarization, segments the input speech signal into clusters based on speaker occurrence.

The need for speaker segmentation in HAAWAII is to pre-process large amounts of data provided by ANSPs. Machine learning models require large amounts of data for training. In case of HAAWAII project, this data is delivered by ANSPs in a raw format, i.e., not segmented to speech utterances by SAD and often combining both speaker channels (ATCo and pilots) in the same file. To prepare good training data, typically the speech utterances (few seconds long) obtained by the SAD module are then segmented according to speaker by the speaker segmentation. As hitherto mentioned, the speaker segmentation is required as an offline module to substitute the PTT which is often not available by ANSPs.

a) Evaluated speaker segmentation technologies: We applied a speaker diarization system as described by Landini et al. [12]. It is based on the clustering of speaker embeddings — "x-vectors". The x-vectors are obtained from a neural network trained to discriminate speakers so the embeddings capture relevant information that allows comparing them and deciding when two of them correspond to the same person. For clustering, a Bayesian hidden Markov model is used where each state represents one speaker. When finding the state that most likely is produced by a given x-vector, the x-vector is assigned one speaker. Eventually, defining the final assignment of x-vectors to speakers denotes the diarization output.

The assignment of x-vectors to speakers thus defines a part of the recording where the given speaker is speaking. Generally, in each recording, many speakers can be identified as there appear several pilots and supposedly one ATCo. According to this prior assumption, the speaker with the most speech parts in a recording is marked as a controller (ATCo), and all other speakers are marked as pilots. Note that this system was used to pre-process the audio files for annotators whose task is to further verify/improve the audio pre-segmentation and possibly manually correct the speaker identity.

[5]https://catalog.ldc.upenn.edu/LDC2019S22

TABLE I. EXAMPLE OF ATC COMMUNICATION FOR EACH SPEAKER ROLE I.E., ATCO AND PILOT

Speaker Label	Transcript
ATCo	$< s >$ two echo golf taxi to holding point delta runway two five $< /s >$ $< s >$ skytravel two seven eight six proceed to rudap $< /s >$
Pilot	$< s >$ heading three six zero degrees speed bird four seven five $< /s >$ $< s >$ london hello speed bird four one three flight level one three zero $< /s >$

In HAAWAII, two approaches were analyzed and are described in this paper, both functioning on very different data types: acoustic and text-based. Further in this paper, the speaker diarization task will be referred to as the speaker role detection task, for small segments of audio which are more briefly described below.

C. Speaker Role Detection

For a given speech or text segment, the aim of this module is to reliably detect the role of the speaker – ATCo or pilot. This module is necessary when no PTT signal is available and the inbound and outbound communication is merged into a channel. Speaker role detector is applied on top of SAD and speaker segmentation (i.e., these modules are able to segment the input speech into short utterances and cluster them to 2 speaker classes, but we cannot directly predict which class corresponds to ATCo or pilot. To accomplish this, we experimented with acoustic based and text-based approaches in HAAWAII.

a) Acoustic based approach: In speaker recognition, I-vectors [13], and more recently x-vectors [14], are among the most commonly employed representations. Current hybrid ASR systems use online i-vectors as input features to make the acoustic model robust to the speaker variability [15], [16]. The i-vectors are computed frequently (e.g., every 10 seconds), which especially helps to reduce the latency in online decoders. However, in speaker recognition systems, as shown in [14], the x-vector significantly outperforms standard i-vector systems as it exploits the modeling power of deep neural networks. HAAWAII investigated the integration of i-vector and x-vector speaker models at the speech utterance level to detect the role of the speaker in each utterance provided by SAD.

Similar to [14], the i-vector and x-vector systems trained on SRE16 data are used and the corresponding Probabilistic Linear Discriminant Analysis (PLDA) [17] classifier is then adapted with publicly available ATC communication data (LDC-ATCC [18] and UWB-ATCC[6]). These datasets span

[6]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-CCA1-0

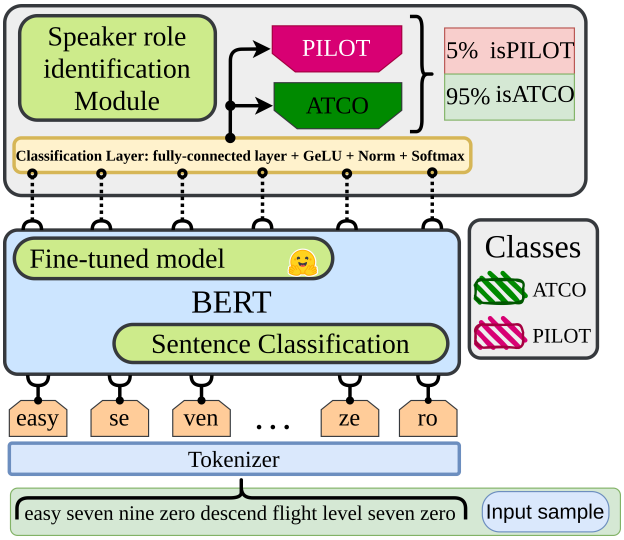

Figure 2. BERT-based speaker role identification module.

~40 h of speech data related to both phraseology and structure seen in ATCo-Pilot communication out of which 33 h (18 h and 15 h for ATCo and piot respectively) is used for training. This data is additionally augmented by adding noise using the open-source Musan [19].

A text-based processing method can be seen as a way to correct wrongly assigned speaker identities by the acoustic method. It takes advantage of the availability of ASR transcripts and is able to achieve relatively high accuracy. On the other hand, the module operates on top of already generated segments (i.e., speech utterances) and is not capable of changing the segment boundary or splitting the segment if there are two (or more) speakers present.

b) Text-based approaches: The following two text-based processing approaches were studied and implemented in HAAWAII: **Rule based -** ATC communication provides rich source of information that follows explicit grammar and ontology. Additionally, ATC communication (see examples in Table I) exploits a well-defined lexicon and dictionary, even though sometimes disrupted by speakers' errors. One example is the order in which named entities (e.g., callsign) are uttered during the communication; ATCos utter callsigns at the beginning, while pilots constantly at the end.

ICAO defines a phraseology to be followed by ATCos and pilots to enable clear communication. Phraseology elements used by pilots obviously distinguish from ATCo elements. For instance, there are certain phrases that an ATCo should use in a specific order. This knowledge can therefore be used to extract/identify potential words/phrases that further indicate a specific role of the speaker as shown in [20].

Data-driven - A substantial effort in the rule-based approach is creating a so called "bag of words" for both classes. The words have to be carefully extracted, and feedback from an expert is required, which demands costly manual involvement. In order to improve this approach, various rules have to be incorporated (e.g., using all variants of a callsign, or using the words before the callsign is uttered), which leads to the complex system impossible to be deployed in real-time. As an alternative, a data-driven approach can be used, which learns the information from data to assign a class to each phrase. The following two approaches were developed:

1. Convolution Neural Network (CNN)-based [21] detector: Each sentence is represented by 96-dimensional word embeddings, which are passed through 64 x 4 convolution filters, each of width 1, 2, 3, and 4. Max pooling on the resulting feature maps was used, which resulted in a fixed dimensional sentence representation. This is passed through a two-layer feed-forward network with ReLU activations and finally a softmax output function for controller/pilot class-probabilities. All the model parameters (including word embeddings) are updated together. In total, the model has about 430K trainable parameters.

2. Bidirectional Encoder Representations from Transformers (BERT)-based detector [22]: Sequence Labeling (SL) assigns labels to words that share a specific role and meaning within the grammatical structure of a sentence. In [23], these groups of words/sentences have similar grammatical properties. Their work focuses on two sub-tasks of SL: Named Entity Recognition (NER) [24], [25] and Sequence Classification (SC) [23], [26]. Early work on NER and SC was based on handcrafted ontology, dictionaries, and lexicons, making them prone to human errors. Since the past decade, deep learning based systems have been cataloged as state-of-the-art on NER [27] and SC. These models are mostly based on convolutional neural networks [21], recurrent neural networks [23], and transformers [22].

Following the aforementioned pros and cons, we believe the state-of-the-art NER and SC can be leveraged to identify speaker roles. For instance, one can apply NER to identify ATC-related named entities such as *callsigns*, *commands*, or *units*. Similarly, the structure and the type of these 'entities' used in a given communication can be leveraged to identify speaker roles. Our previous research on identifying speaker roles [20] was mainly directed as a grammar-based bag of words system that was capable of performing speaker role identification. In [28], the authors mention that even manually annotating pilot recordings is twice as hard compared to ATCo recordings due to their quality, rate of speech, speaker accent, etc. This is one of the reasons why speech processing systems (ASR, diarization, and speaker role identification) perform considerably worse on pilots than ATCos' recordings.

We implemented a BERT-based speaker role identification module (see figure 3) that allows to attribute a speaker role (i.e., ATCo or pilot) to a given ATC speech utterance. We fetched a BERT [22] from Huggingface [29]. We then use ground truth transcripts to fine-tune the model on the sequence classification task with HAAWAII data, which is defined in Table II. Around 15K sentences of both ATCo and pilot class were used for fine-tuning. The transformer (BERT) model is fine-tuned for 5 epochs (weight decay set to 0.01, and 500 warm-up steps). After the fine-tuning, we perform inference

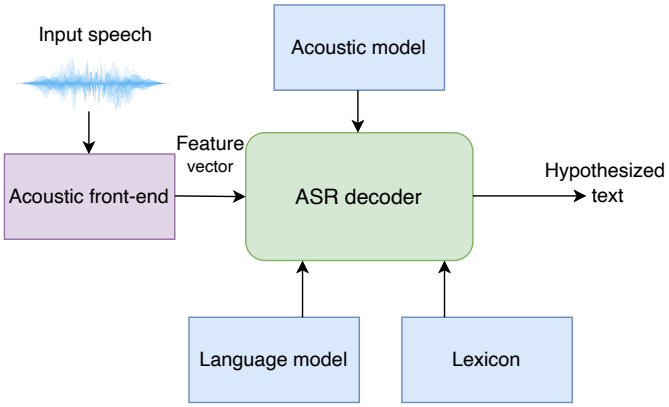

Figure 3. Building blocks of a typical ASR system combining language and acoustic models.

on either manually transcribed or automatically generated ATC speech utterances.

D. Automatic Speech Recognition

Automatic speech recognition (ASR) or speech-to-text systems convert speech to text – the system receives the input signal S (usually segmented by SAD to short speech utterances) and transforms it into a sequence of words W:

$$\hat{W} = \underset{W \in \mathcal{V}^*}{\operatorname{argmax}}\, p(W|S),$$

where $\mathcal{V}$ is the vocabulary of all possible words; we use $\mathcal{V}^*$ to represent the collection of all word sequences formed by words in $\mathcal{V}$.

Two main blocks of an ASR system are the Acoustic Model (AM) and the Language Model (LM). The AM represents the relationship between the speech signal and the phonemes/linguistic units that make up speech and is trained using speech recordings along with their corresponding text transcripts. The LM provides a probability distribution over a sequence of words, provides context to distinguish between words and phrases that sound similar, and is trained using a large corpus of text data. The AM and the LM are eventually combined in the following way:

$$\begin{aligned} \hat{W} &= \underset{W \in \mathcal{V}^*}{\operatorname{argmax}}\, p(W|S) && (1)\\ &= \underset{W \in \mathcal{V}^*}{\operatorname{argmax}}\, p(S|W)p(W) && (2)\\ &= \underset{W \in \mathcal{V}^*}{\operatorname{argmax}} \sum_{P} p(S|P)p(P|W)p(W), && (3) \end{aligned}$$

where $p(S|P)$ represents the AM, $p(P|W)$ is a pronunciation model, and $p(W)$ is the LM. Specifically in HAAWAII, a hybrid ASR engine is deployed. The AM is represented by a neural network while the LM is n-gram model.

Once the ASR engine is built (both AM and LM are sufficiently trained), the recognition is performed by decoding the input speech. To do so, a decoding graph is typically used, which combines different ASR building blocks together and is represented as a weighted finite state transducer (WFST) [**?**], [30], [31]. Given an input speech utterance, word recognition 'lattices' can be generated during the decoding, which contain the most likely hypothesised word sequences.

E. Boosting of important word entities for ASR

To enhance the accuracy of hypothesised ASR output, additional contextual information can be incorporated alongside the spoken input. To use the contextual information with the hybrid ASR, a technique, known as lattice rescoring with WFST, was previously proposed to bias the system towards users' playlists [32], contact names [33], and named entities [34].

In the context of HAAWAII, the additional information is presented by data obtained from radar systems. Radar continuously tracks aircraft within a given airspace, providing unique identifiers known as 'callsigns' for each aircraft. A callsign typically comprises a sequence consisting of an ICAO airline identifier, letters, and digits, which are spoken as a sequence of words. Utilizing this radar data allows to identify the callsigns that are most likely to be mentioned in the conversation. This knowledge enables us to adjust the ASR system's predictions in favor of these registered callsigns, increasing the likelihood of their correct recognition.

During the ASR decoding, the desired word sequences (to be boosted) can be given more importance by utilizing WFST and adjusting weights within the prediction graphs, i.e. 'lattices.' Recently, a similar biasing approach has shown promise in improving callsign recognition [35]–[37]. Biasing the lattices with context-related callsigns has consistently shown significant improvements in their recognition within the final output. Therefore, lattice rescoring was applied to enhance the callsign recognition in the HAAWAII framework.

F. Adaptation of ASR to New Environment

The goal of adaptation is to re-use a (seed) model trained with well-resourced ATC communication data, and further fine-tune it with the under-resourced (i.e., target) domain scuh as new airport/airspace). In HAAWAII, as seed model, we use model trained on all ATC available (manually transcribed) speech data and adapt it to the target airport. (i.e., NATS). To better understand the model adaptation capabilities, the amount of data from the target airport vary. We consider both models to be adapted:

- AM adaptation: This entails fine-tuning the neural acoustic model trained using mixed data to the under-resourced target domain. The goal here is to fine-tune the model parameters to the acoustic conditions and speaker/pronunciation variations common in the target airport/airspace.
- LM adaptation: This approach updates the n-gram LM probabilities pre-computed with the available ATC data towards the target domain.

TABLE II. HAAWAII TRAIN AND TEST SET CHARACTERISTICS. † - SPEAKER ROLE DETECTION WAS APPLIED TO SPLIT THE DATA ACCORDING TO SPEAKER ROLES, ATCO / PILOT.

Dataset	Nb. utts [x 10^3]†	Dur [h]†
	Train set	
Isavia	7.2 / 8.4	9 / 10
NATS	11.5 / 12.6	11.8 / 12.3
	Test set	
Isavia	0.5 / 0.6	0.5 / 0.6
NATS	0.4 / 0.5	0.4 / 0.4

TABLE III. COMPARISON OF SAD RESULTS ON EXTERNAL LIVEATC EVALUATION SET DESCRIBED IN SECTION III-A.

SAD model	**DetER(%)**	FA(%)	Miss (%)
ASR SingleBest	10.1	4.9	5.2
ASR Mul LR	**9.7**	6.1	**3.6**
ASR Mul MV	11.1	**4.3**	6.8
Phn Rec	20.1	4.6	15.5
WebRTC	16.5	9.4	7.1
Pyannote	13.8	10.1	3.7
Energy-based	13.3	7.1	6.2

III. PERFORMED EXPERIMENTS

This section describes experiments and presents objectively obtained performance analyses for the modules introduced in previous section II. The individual speech processing modules are evaluated on the following datasets: (i) London TMA data collected from NATS ANSP (referred to as NATS) and (ii) Iceland enroute set collected by Isavia ANSP (referred to as Isavia). These data were collected and manually transcribed and annotated as part of HAAWAII project. The distribution of training and evaluation subsets of NATS and Isavia is presented in Table II.

A. SAD

Different SAD approached, introduced above, are evaluated using the following metrics: False Alarm (FA) rate, Miss detection (Miss) rate, and Detection Error Rate (DetER). FA presents the number of non-speech utterances being falsely detected as speech, while Miss detection rate presented the number of speech utterances mis-detected by SAD. Specifically for $DetER$, it is defined as:

$$\textbf{DetER} = \frac{FA\,(s) + Miss\,(s)}{Total\,duration\,of\,speech\ (s)},$$

where s means the length of speech utterance in seconds.

Motivated by [7], we compared the following SAD approaches on external LiveATC data. LiveATC are ATC communication data collected in an automatic manner from VHF channels as part of ATCO2 project[7]. The total duration of the data is 6.8 h, consisting in total 1000 speech utterances.

[7]https://www.atco2.org

TABLE IV. SUMMARY OF SPEAKER SEGMENTATION PROCESS ON ISAVIA AND NATS DATA (PRESENTED IN (%)).

Dataset	**# utterances**	Insertion	Deletion	Splits	Speaker change
Isavia	3225	0.00	9.4	0.3	5.9
NATS	6235	0.01	2.4	0.05	1.1

A comparison of SAD results on the LiveATC evaluation set is presented in Table III. It needs to be mentioned that the neural based approaches such as 'ASR SingleBest', 'ASR Mul MV' and 'ASR Mul LR' do not use any ATC data for training respective models. All ASR based models (the ASR Mul LR model, ASR SingleBest and ASR Mul MV models) significantly outperformed the baseline production models such as WebRTC or Pyannote. Interestingly Energy-based SAD (also applied for real-time processing in HAAWAII) yields good performance.

B. Speaker Segmentation

Performance of speaker segmentation on ATC data is presented in Table IV. Specifically we show how many times the speaker segmented output was modified by annotators. Following numbers are presented:

- Insertions - the number of speech utterances newly added by annotators,
- Deletions – deleted speech utterances are mostly concatenated with preceding/following ones,
- Splits - an utterance was divided into two, and
- Speaker change - the number of times the speaker was incorrectly classified, and the speaker label was corrected.

It can be observed that the deletion rate for Isavia is significantly higher. One of possible explanation is that the automatically created segments are very short and thus were concatenated with previous/following segments.

C. Speaker Role Detection

The different text-based speaker role detection systems are evaluated using the accuracy metric defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN},$$

where TP is the number of times the system correctly recognizes the ATCo, TN is the number of times the system correctly recognizes the Pilot, FN is the number of times the system incorrectly recognizes ATCo as Pilot and FP is the number of times the system incorrectly recognizes Pilot as ATCo.

Table V shows the performance of the proposed models on the ISAVIA and NATS test sets. The results reveal that the text-based approach outperforms the acoustic based x-vector system. Among the text-based systems, the BERT-based model outperformed all other systems across different test sets.

TABLE V. ACCURACY (%) OF ACOUSTIC AND TEXT-BASED SPEAKER ROLE DETECTION SYSTEMS FOR ISAVIA AND NATS TEST SETS (SEE TABLE II) FOR DATA DESCRIPTION.

Model	Isavia	NATS
x-vector plda	76.5	83.8
Rule-based	82.0	85.0
CNN	91.0	93.0
BERT	**93.0**	**96.0**

D. ASR

Throughout ASR development and tests, Kaldi framework [38] was applied. The developed ASR is built around standard Kaldi recipe which uses MFCC and i-vectors features. The standard AM training is based on Lattice-free MMI (LF-MMI) [11], which includes 3-fold speed perturbation and one-third frame sub-sampling.

The AM uses a conventional biphone convolutional neural network (CNN) [21] + Factorized Time Delay Neural Network (TDNN-F) [39] model while the LM is a statistical 3-gram trained on the same data as the acoustic model with additional textual data collected from public resources such as airline names, airports, the ICAO alphabet, and way-points in Europe. Specifically for training,

- Baseline ASR: trained with 100 h of transcribed ATC data (further augmented with speed perturbation, obtaining 300 h) which does not include any HAAWAII data. Specifically LDC-ATCC, UWB-ATCC, ATCOSIM, and MALORCA sets described in [40] are used.
- HAAWAII ASR: trained with approximately 195 hours of ATC manually transcribed data [40] (further augmented using the speed perturbation, obtaining almost 575 hours of training data).

The performance of an ASR system is presented in terms of Word Error Rate (WER). It is based on the Levenshtein distance at the word level and it can be viewed as a string matching problem where two sequences of symbols are matched through the dynamic programming. The symbols in this case are the words of a language. WER finds the distance between the word sequence hypothesised by the ASR and the reference word sequence using dynamic string alignment i.e., it finds the number of edits (substitutions, deletions, insertions) required to go from the hypothesised word sequence to the reference word sequence. In other words, given the hypothesised and reference sequences of words, WER is computed as:

$$WER = \frac{S + D + I}{N}, \tag{4}$$

where S is the number of words that are substituted, D is the number of deletions, I is the number of insertions, $N = S + D + C$ is the total number of words in the reference and C is the number of correctly recognized words. A lower WER implies higher accuracy for the ASR system.

Table VI shows the WER of the final hybrid LF-MMI based ASR developed using HAAWAII data. The results yield

TABLE VI. WER (%) ASR RESULTS FOR NATS AND ISAVIA TEST SETS.

ASR system	WER(%)	
	Isavia	NATS
Baseline ASR	-	28.3
HAAWAII ASR	12.4	7.5

TABLE VII. RESULTS FOR BOOSTING CALLSIGNS ON ISAVIA AND NATS TEST SETS. ¶WORD ERROR RATES (ENTWER) ESTIMATED ONLY FOR THE CALLSIGN UTTERANCE.

Boosting	Isavia			NATS		
	WER	EntWER¶	ACC	WER	EntWER¶	ACC
HAAWAII ASR	12.4	5.0	87.9	7.5	4.1	86.7
Unigrams	12.3	3.7	90.7	7.4	3.6	88.0
N-grams	12.1	4.1	90.5	6.7	2.0	93.3
GT boosted	11.6	2.5	94.7	6.4	1.3	96.1

WER of 7.5% and 12.4% for NATS and Isavia respectively (i.e., 28.3% WER obtained for the baseline ASR trained on other ATC data). We observe an absolute difference of 5% in WER between Isavia and NATS due to the different acoustic conditions, and accents in the former.

E. Boosting of important word entities for ASR

Table VII shows the results on boosting experiments reporting the WER of the whole speech utterance, WER estimated on the callsigns only (EntWER), and the accuracy of correctly recognizing callsign (corrent/incorrect) for Isavia and NATS test sets. The baseline is represented by the HAAWAII ASR not applying any boosting (biasing). Other three types of experiments apply callsign boosting and differ from each other by how and which callsigns are specifically boosted in final ASR lattices. *Unigrams* boosting means biasing towards only single words which are taken from the callsigns registered in the surveillance data. *N-grams* boosting means that all callsigns in the surveillance data from a current time stamp are boosted as word sequences. In order to present an 'oracle' performance for the biasing method, the ASR lattice is biased only toward a single *ground truth (GT)* callsign, and the boosting is done as for a word sequence.

Biasing the lattice toward the context callsigns usually allows us to considerably improve their recognition in the final outputs. Various experiments conducted on ATC data have consistently shown that employing lattice rescoring on top of ASR predictions results in higher accuracy for automatic transcriptions, particularly for callsigns [41].

F. ASR adaptation to new environment

Table VIII yields the ASR performance when developed ASR is adapted to the target airport. We use the baseline ASR described in Section III-D as seed model to be adapted to the target airport – NATS. Experiments are conducted for different amounts of transcribed data obtained from NATS. The baseline

ASR system when adapted to the target NATS airport with 14h of transcribed data provides a WER of 13.9%.

TABLE VIII. WER (%) FOR NATS TEST FOR VARIOUS ADAPTATION SETTINGS USING A GENERIC ATC ASR MODEL

Target	**Type of adaptation WER(%)**			
data (h)	Baseline ASR	AM	LM	AM+LM
0	28.3	-	-	-
1	-	23.3	23.3	23.0
2	-	22.3	21.2	22.5
4	-	19.7	19.2	18.4
10	-	17.6	17.3	15.6
13	-	16.1	17.3	**13.9**

IV. CONCLUSION

The HAAWAII project presents an innovative framework to recognize and understand the air-traffic communication. This paper described different components of the whole framework, focusing on employed speech technologies. The paper presents approaches for both offline and real-time speech processing. Specifically speech activity detection, speaker segmentation and role detection, and automatic speech recognition modules integrated in the HAAWAII framework were described. Eventually a technique for rapid adaptation of ASR engine to the target environment (in this case presented by NATS airport) was presented.

ACKNOWLEDGMENT

The work was supported by SESAR Joint Undertaking under European Union's Horizon 2020 research and innovation programme under EC project No. 884287 - HAAWAII.

REFERENCES

[1] H. Helmke, M. Kleinert, N. Ahrenhold, H. Ehr, T. Mühlhausen, O. Ohneiser, L. Klamert, P. Motlicek, A. Prasad, J. Zuluaga-Gomez *et al.*, "Automatic speech recognition and understanding for radar label maintenance support increases safety and reduces air traffic controllers' workload," in *Fifteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)*, 2023.

[2] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. 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.

[3] M. Kleinert, H. Helmke *et al.*, "Semi-supervised adaptation of assistant based speech recognition models for different approach areas," in *37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018, pp. 1–10.

[4] 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, pp. 1–10.

[5] 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.

[6] H. Helmke, M. Kleinert *et al.*, "The HAAWAII Framework for Automatic Speech Understanding of Air Traffic Communication," in *submitted to 13th SESAR Innovation Days, Seville, Spain*, November 2023.

[7] S. S. Sarfjoo, S. Madikeri, and P. Motlicek, "Speech Activity Detection Based on Multilingual Speech Recognition System," in *Proc. Interspeech 2021*, 2021, pp. 4369–4373.

[8] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[9] P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, T. Mikolov, and O. Plchot, "BUT system description for NIST LRE 2007," in *Proc. 2007 NIST Language Recognition Evaluation Workshop*, 2007, pp. 1–5.

[10] S. Madikeri, B. K. Khonglah, S. Tong, P. Motlicek, H. Bourlard, and D. Povey, "Lattice-free maximum mutual information training of multilingual speech recognition systems," in *Proc. of Interspeech*, vol. 2020, 2020.

[11] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[12] 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.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[15] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[16] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[17] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

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

[19] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[20] A. Prasad, Z.-G. Juan, P. Motlicek *et al.*, "Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition," in *12th SESAR Innovation Days*, 2022.

[21] 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, p. 1995, 1995.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] Z. He, Z. Wang *et al.*, "A Survey on Recent Advances in Sequence Labeling from Deep Learning Models," *arXiv preprint arXiv:2011.06727*, 2020.

[24] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[25] V. Yadav and S. Bethard, "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2145–2158.

[26] C. Zhou, B. Cule, and B. Goethals, "Pattern based sequence classification," *IEEE Transactions on knowledge and Data Engineering*, vol. 28, no. 5, pp. 1285–1298, 2015.

[27] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[28] 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," *arXiv preprint arXiv:1810.12614*, 2018.

[29] T. Wolf *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*. Association for Computational Linguistics, 2020, pp. 38–45.

[30] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.
[31] M. Riley, C. Allauzen, and M. Jansche, "OpenFST: An open-source, weighted finite-state transducer library and its applications to speech and language," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, 2009, pp. 9–10.
[32] K. Hall, E. Cho, C. Allauzen, F. Beaufays, N. Coccaro, K. Nakajima, M. Riley, B. Roark, D. Rybach, and L. Zhang, "Composition-based on-the-fly rescoring for salient n-gram biasing," 2015.
[33] P. Aleksic, M. Ghodsi, A. Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach, and P. Moreno, "Bringing contextual information to google speech recognition," 2015.
[34] J. Serrino, L. Velikovich, P. S. Aleksic, and C. Allauzen, "Contextual recovery of out-of-lattice named entities in automatic speech recognition." in *Interspeech*, 2019, pp. 3830–3834.
[35] M. Kocour, K. Veselý, I. Szöke, S. Kesiraju, J. Zuluaga-Gomez, A. Blatt, A. Prasad, I. Nigmatulina, P. Motlíček, D. Klakow *et al.*, "Automatic processing pipeline for collecting and annotating air-traffic voice communication data," *Engineering Proceedings*, vol. 13, no. 1, p. 8, 2021.
[36] I. Nigmatulina *et al.*, "Improving callsign recognition with air-surveillance data in air-traffic communication," Idiap Research Institute. Idiap Research Institute, 2021, pp. 1–5.
[37] I. Nigmatulina, J. Zuluaga-Gomez, A. Prasad, S. S. Sarfjoo, and P. Motlicek, "A two-step approach to leverage contextual data: speech recognition in air-traffic communications," in *ICASSP*, 2022.
[38] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.
[39] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.
[40] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina, P. Motlicek, and M. Kleinert, "A Virtual Simulation-Pilot Agent for Training of Air Traffic Controllers," *Aerospace*, vol. 10, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2226-4310/10/5/490
[41] J. Zuluaga-Gomez *et al.*, "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.