BERTraffic: BERT-based Joint Speaker Role and Speaker Change Detection for Air Traffic Control Communications

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

[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[3] Brno University of Technology, Brno, Czech Republic
[4]German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany

juan-pablo.zuluaga@idiap.ch

Abstract

Automatic speech recognition (ASR) allows transcribing the communications between air traffic controllers (ATCOs) and aircraft pilots. The transcriptions are used later to extract ATC named entities e.g., aircraft callsigns, command types, or values. One common challenge is Speech Activity Detection (SAD) and diarization system. If one of them fails then two or more single speaker segments remain in the same recording, jeopardizing the overall system's performance. We propose a system that combines the segmentation of a SAD module with a BERT model that performs speaker change detection (SCD) and speaker role detection (SRD) by chunking ASR transcripts i.e., diarization with a defined number of speakers together with SRD. The proposed model is evaluated on real-life ATC test sets. It reaches up to ∼0.90/∼0.95 F1-score on ATCO/pilot SRD, which means a 27% relative improvement on diarization error rate (DER) compared to standard acoustic-based diarization. Results are measured on ASR transcripts of challenging ATC test sets with ∼13% word error rate, and the robustness of the system is even validated on noisy ASR transcripts.

Index Terms: Text-based diarization, speaker change detection, speaker role detection, air traffic control communications, chunking.

1. Introduction

Air traffic controllers (ATCOs) supervise a portion of airspace by issuing commands to pilots. Most of these voice-based communications are conveyed over noisy VHF (very-high frequency) channels i.e., low signal-to-noise ratio (SNR). In a typical scenario, the ATCO (speaker1) issues voice-based commands to pilots (speaker2) together with pre-defined callsigns (name of the flight). Considering that a big portion of this communication is transmitted via voice messages, previous studies proposed to apply ASR to automatically extract the corresponding transcripts. In recent years, the ASR systems were shown to reach a maturity in reducing ATCO's workload, but for research-only scenarios. Examples are AcListant®-Strips [1] and MALORCA[1] projects. The later shows that novel data-driven machine learning approaches enable costly adaptations to different airport environments [2]. Lin [3] reviews ten tasks on spoken instruction understanding of ATC data. Semi-supervised learning has been also explored on the framework of ATC [4]. Ongoing HAAWAII[2] and SOL-Cnt projects focus on developing a reliable and adaptable solution to automatically transcribe voice commands issued by both ATCOs and pilots. For instance, higher accent variability and noise level cause ASR systems by factor of two the word error rates (WER) for pilots' compared to ATCOs' [5] speech. In addition, closeness and overlap (cross-talk) between speech segments of ATCO and pilots cause acoustic-based diarization systems to yield non-acceptable levels of diarization error rate (DER), jeopardizing the SCD step and subsequently the ASR system end up processing utterances with multiple speakers.

Table 1: *Conversation between two speakers with correct SAD and SCD (rows 1 and 2) and SCD fault (row 3).* [†]*samples from SOL-Cnt test set.*

Speaker Label	Detected segment[†]
ATCO (speaker 1)	<s> november six two nine charlie tango report when established </s>
Pilot (speaker 2)	<s> report when established november six two nine charlie tango </s>
Mixed (SAD and SCD failed)	<s> november six two nine charlie tango report when established report when established </s> <s> november six two nine charlie tango </s>

1.1. Motivation and contribution

In this work, we fine-tune a pre-trained BERT model to jointly perform tagging and chunking for SCD and SRD. Chunking allows to split sentences into tokens (or words) and then merging them in a meaningful sub-groups. In our case, a phrase (or entity) is composed of a full single-speaker ATC utterance, with either ATCO or pilot as speaker role (see Table 1). By applying chunking in a multi-speaker and multi-segment (or single-speaker and single-segment) utterance, one can perform speaker change detection (SCD) and speaker role detection (SRD) simultaneously on the text level (Figure 1 mid-box). Our approach is tested on five single-speaker per utterance test sets and one multi-speaker test set (i.e., SAD failure). We also developed a simple but effective data augmentation pipeline to

⋆ equal contribution.
† corresponding author.

This work was supported by the SESAR Joint Undertaking under Grant Agreement No. 884287, under European Union's Horizon 2020 Research and Innovation programme.

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

[2]https://www.haawaii.de

counteract the class imbalance within the train sets. Text-based diarization (i.e, combination of SAD, SRD, and SCD) yielded 0.90/0.95 F1-scores on SRD (ATCO/pilot) in single-speaker test sets while an 0.89 F1-score in the multi-speaker one. We also compare text-based with conventional acoustic-based diarization. Experiments are conducted on both, ground truth annotations and transcripts generated by our state-of-the-art hybrid-based ASR system for ATC [6].

2. Related Work

Diarization systems answer the question *"who spoke when?"*. SAD, segmentation or SCD, embedding extraction, clustering and labeling are the main parts of a diarization system.

Acoustic-based diarization: feature representations of speakers are one of the main factors in the accuracy of a speaker diarization system. Mel frequency cepstral coefficients (MFCCs) are commonly used for the task of speaker diarization. In comparison to MFCC, mel filterbank slope (MFS) and linear filterbank slope (LFS) features have more speaker discriminability power caused by emphasis on higher-order formants [7]. Agglomerative information bottleneck (aIB) approach to speaker diarization has shown competitive performance [8]. Here, for clustering the fixed-length audio segments, a bottom-up clustering approach is applied in the posterior space represented by a mixture of gaussians. Speaker discriminative embeddings such as x-vectors are investigated in [9]. For finding the speaker clusters in a sequence of x-vectors, the variational bayesian hidden Markov model (VBx) was investigated in [10]. For continuously learning speaker discriminative information, "Remember-Learn-Transfer" was proposed in [11]. Applying lexical and acoustic information for speaker diarization was investigated in [12].

Text-based speaker role detection: Early text-based techniques relied on handcrafted lexicons, dictionaries, and rules. They were prone to human errors and not robust against noisy labels e.g., produced by standard ASR systems. Collobert et al. [13] introduced machine learning methods for text processing in part-of-speech tagging, chunking, and semantic role labeling. In [14], domain-based chunking is addressed, which is similar to the approach proposed in this paper. The reader might relate chunking to named entity recognition (NER). NER is a chunking sub-task that aims at identifying entities on text e.g., locations, organizations, or names [15, 16, 17]. Examples of named entities in ATC communications are *callsigns*, *command types*, etc. These entities carry rich information that gives insights about the speaker's role (ATCO or pilot). A recent work aligned to ATC domain is reviewed by Prasad et al [18]. Here, a grammar-based approach performs SRD on single-speaker utterances, with the drawback that it does not work on multi-speaker utterances. In [19] a text-based SRD for multiparty dialogues is proposed, but limited to SRD. Finally, text-based diarization has been proposed in the past by [20, 21]. This work leverages tagging and chunking to perform SRD and SCD on ATC text. In general, chunking is used to parse phrases from unstructured text e.g., in our case tagging and chunking a utterance will detect speaker changes and its speaker roles.

3. Datasets and Experimental Setup

This research uses six datasets in the English language with various accents that are collected over four different projects. For all experiments, the validation set is taken directly from the training set.

Table 2: *Amount of train and test data. † real-life ATC set where diarization or speech activity detection failed.*

	Number of Samples (Train/Test)		
Project - dataset	ATCO	Pilot	Mixed
SOL - SOL-Cnt†	662 / 138	945 / 204	535/205
SOL - SOL-Twr	1399 / 594	- / -	-/-
MALORCA - Vienna	6335 / 1557	- / -	-/-
MALORCA - Prague	1364 / 1419	- / -	-/-
HAAWAII - NATS	4228 / 631	4782 / 758	-/-
HAAWAII - ISAVIA	2625 / 493	3046 / 590	-/-

3.1. Datasets

SOL-Cnt & SOL-Twr: two data sets recorded and collected over three SESAR-2020 funded industrial research projects. First, **"SOL-Cnt"** (PJ.10-W2-96) aims to reduce ATCOs' workload with an ASR-supported aircraft radar label. Voice utterances of ATCOs and pilots have been recorded in the operations room at the air navigation service provider (ANSP) site of Austrocontrol in Vienna, Austria. Then, **"SOL-Cnt"** (mix of two projects: PJ.16-W1-04[3] and PJ.05-W2-97[4]), aims to reduce ATCOs' workload in an ATC tower environment (English speech of ATCOs from Lithuania). The SOL-Twr audio data is of better quality compared to SOL-Cnt due to the laboratory recording environment [22].

MALORCA: one of the first initiatives to overcome the need of significant expert knowledge on ASR systems for ATC tasks [4, 23]. During the project two datasets were collected: Prague and Vienna approach (airports), detailed in Table 2. Both datasets only contain ATCO utterances and are sampled at 8k Hz (cataloged as good quality speech i.e., SNR $\geq$ 20 dB).

HAAWAII: speech data is collected and annotated by ANSPs: (i) London approach (NATS) and (ii) Icelandic en-route (ISAVIA). For NATS, the amount of manually transcribed data available is around 10 h (partitioned into 9 h for train and 1 h for test). For ISAVIA, a total of 13 h of manually transcribed data is available (partitioned into 12 h for train and 1 h for test). NATS recordings are of better quality than ISAVIA. Previous work is covered in [18, 24, 25].

3.2. System description

In addition to manual speech transcripts, speaker labels and time segmentation (e.g., ATCO/pilot/mixed) are also available. The BERT model first tags each word of the transcript (ground truth or ASR transcript) with one tag of the *IOB format* (Inside-Outside-Beginning). In IOB format each entity (a full sentence in our case) is composed of two tags, the *Beginning **'B-'*** and the *Inside* tag ***'I-'***. We define ATCO recordings as *Speaker1* (green), while pilot segments as *Speaker2* (red). We do not use the *Outside* tag, because we know that each word is always from one of two predefined speakers.

3.3. Data augmentation

We implemented a data augmentation pipeline to counteract the large class imbalance in the train sets i.e., 64%/33%/3% for ATCO/pilot/mixed recordings (see Table 2). First, we split the

[3]https://www.sesarju.eu/projects/cwphmi

[4]https://www.remote-tower.eu/wp/project-pj05-w2/solution-97-2

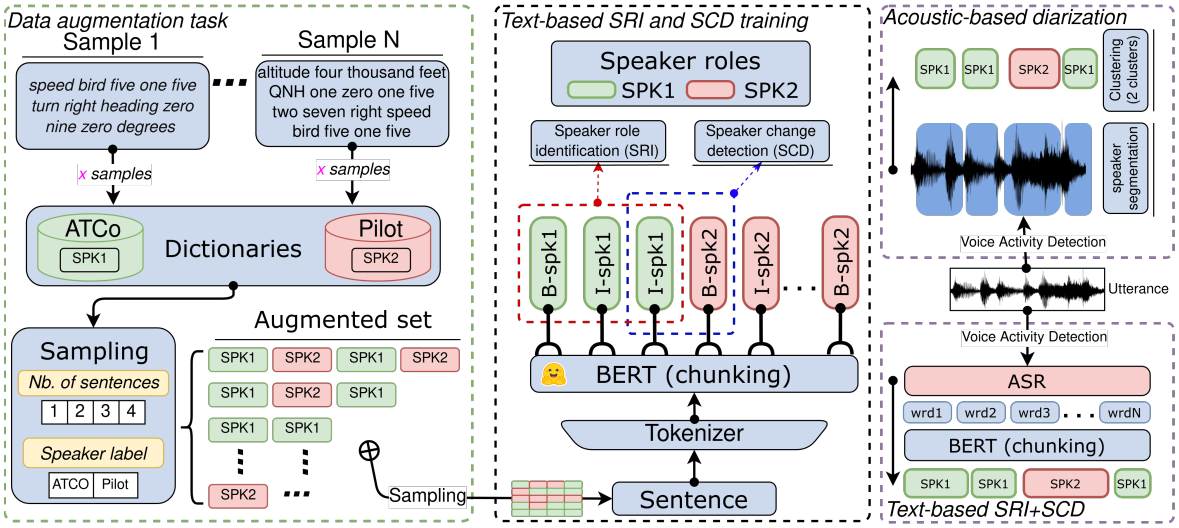

Figure 1: ***Left block:*** *proposed data augmentation pipeline. New samples contain between one to four sentences (probabilities of 40%, 30%, 20% and 10% for one to four sentences, respectively). New sentences have equal chance to be sampled from the ATCO or pilot subset.* ***Central block:*** *proposed pipeline to fine-tune a BERT model that performs tagging and chunking for jointly SCD and SRD.* ***Right block:*** *proposed approach to compare acoustic-based diarization and text-based joint SRD and SCD. Note that our acoustic-based diarization system has a maximum of two clusters because ATC communications only contain two speaker roles (ATCO or pilot).*

training sets on either *ATCO* (speaker 1) or pilot (speaker 2) subset. Then, we generate new sentences from the 26k initial utterances, where each new sample depends on: (i) the number of sentences to be concatenated and (ii) speaker label for each new sentence. In general, a new sample is composed by one to four sentences, each with equal chance of being ATCO or pilot class. The process is repeated until gathering $\sim$350 MB of text data (around one million sentences). Left block in Figure 1 depicts the proposed data augmentation pipeline. Finally, we trained two tagging and chunking systems: applying data augmentation either (i) on HAAWAII train set or, (ii) a mix between SOL, MALORCA and HAAWAII training data. From now on, the former is tagged as ***HAAWAII*** and ***ALL_DATA*** for the later.

3.4. Modules

The performance of our text-based SRD and SCD system is contrasted with a standard acoustic-based diarization system on *SOL-Cnt* test set, which is the only one that contains utterances with two or more speech segments i.e., SAD failure. Both systems employ the same multi-lingual ASR-based SAD module [26] to remove the silence in the recording files.

Speaker role and speaker change detection module: the SRD and SCD systems are built on top of a BERT[5] model [27] downloaded from HuggingFace [28, 29]. The model is later fine-tuned with the augmented data set on the tagging and chunking task (following *IOB* format). Each model is fine-tuned for 3k steps ($\sim$5 epochs), with a 500-step warm-up phase. Learning rate is increased linearly until $5e-5$ during the warm-up, then it linearly decays. We fine-tune each model with Adam optimizer, batch size of 32 and, gradient accumulation of 2.

Acoustic-based diarization: we performed aIB-based speaker diarization. For details of the algorithm, the reader is referred to [8]. We used the open-source IB toolkit[6] in our experiment. 19-dim MFCCs were extracted from 8k Hz audio data to match the Gaussian mixture model (GMM) components for the fixed duration segments. The values for normalized mutual information (NMI), the maximum number of clusters, and β were set to 0.4, 2, and 10, respectively. The maximum number of clusters was set based on the maximum number of speakers after applying the SAD.

Automatic speech recognition: a state-of-art hybrid-based ASR system for ATC speech was developed with Kaldi toolkit [30]. The system follows the standard recipe e.g., uses MFCC and i-vectors features with standard chain training based on lattice-free MMI. We use the same ASR system for both ATCO and pilot recordings. Further information about the training recipe and databases is covered in [6, 31, 32, 33].

3.5. Evaluation protocol

The experiments are prepared to answer three questions: (i) How reliable is the BERT-based SRD system on ground truth transcripts? (ii) How is the performance impacted when using automatically generated (ASR) transcripts instead of ground truth transcripts (real-life scenario)? And, (iii) which system performs better on real-life ATC speech data - text-based or acoustic-based diarization?

Speaker role detection: we evaluate SRD with F1-scores on five test sets that are perfectly segmented and we assume only one speaker per segment i.e., one *'B-spk1/2'* tag per utterance. Thus, the task of the system is to detect the correct speaker role. Additionally, the BERT model can tag and chunk sentences with two or more segments from the same speaker role e.g., an utterance with 2 speakers will end up with two segments. Results are shortlisted in Table 3.

Speaker change detection: in addition to SRD, the BERT system performs SCD i.e., central block in Figure 1. We evaluated this task with F1-scores but only on *SOL-Cnt test set*, as it has a subset where SAD failed i.e., multi-segments utterances (**MIXED** column in Table 4). Here, we assume that each utterance contains one or more segments i.e., possibly speaker

[5] *BERT base uncased* model: 110M parameters
[6] https://github.com/idiap/IBDiarization

Table 3: *F1-scores [0-1] in speaker role detection evaluated only on single-speaker utterances.* †*see Section 3.3.* ††*grammar-based SRD system evaluated on both, NATS and ISAVIA [18]. Best performance in* ***bold****.*

	Training Data	
	HAAWAII†	**ALL_DATA**†
Project - Test set	ATCO / PILOT	ATCO / PILOT
Baseline-from [18]††	0.83 / 0.84	- / -
HAAWAII - NATS	0.94 / 0.91	**0.96 / 0.91**
HAAWAII - ISAVIA	0.94 / 0.89	**0.95 / 0.89**
SOL - SOL-Twr	0.82 / -	**0.97** / -
MALORCA - VIENNA	0.81 / -	**0.95** / -
MALORCA - PRAGUE	0.83 / -	**0.95** / -

Table 4: *F1-scores [0-1] in speaker role detection (SRD) and right word boundary (SCD) on SOL-Cnt test set. Results listed for ground truth and ASR transcripts. Best performance is highlighted in* ***bold****.* †*see Section 3.3.* ††*estimated 13% WER.*

Training Data	ATCO	PILOT	MIXED
HAAWAII†			
Ground Truth	0.85	0.87	0.72
ASR output††	0.83	0.84	0.70
ALL_DATA†			
Ground Truth	**0.96**	**0.87**	**0.89**
ASR output††	0.94	0.85	0.82

changes. Table 4 reports our main findings. The F1-scores for SRD and SCD are only reported on the tag *'B-spk1* or *'B-spk2*.

Acoustic-based diarization: to score acoustic-based diarization, we use DER and Jaccard Error Rate (JER) as metrics. DER measures the fraction of time that the segment is not attributed correctly to a speaker or to non-speech. JER is a recently proposed metric [34] that avoids the bias towards the dominant speaker i.e., evaluating equally all speakers. Results are given in Table 5.

4. Results and Discussion

How text-based SRD performs on single-speaker utterances? Results are shortlisted while using two BERT models: (i) trained on HAAWAII train set and we evaluate the robustness of the system on out-of-domain data (SOL and MALORCA test sets); and (ii) trained on all available data (**ALL_DATA** in Table 3 and 4). Initially, we noted that using in-domain data (**ALL_DATA** instead **HAAWAII**) brought 15%, 14% and 12% relative improvement in F1-score for SRD on *SOL-Twr*, *Vienna* and *Prague*, respectively. F1-score in SRD on HAAWAII test set did not increase significantly ($\sim$1%), because domain data was part of the initial training (see Table 3). We noted that both models perform better on ATCO than pilot segments, because pilot's communications include more grammatical errors [18].

Robustness of SCD on ASR transcripts: we evaluated the SCD task only on *SOL-Cnt test set* which comprises recordings with more than one speaker (*mixed* subset). The BERT system is fed with the 1-best transcript obtained from our in-domain hybrid-based ASR system. Table 4 highlights the main results for the two proposed BERT models with an additional line for *'ASR output'*. In the single-speaker case (ATCO/pilot), the degradation (ASR transcripts instead of ground truth text) in SCD from BERT-based diarization was no more than 3% absolute (worse, Pilot subset 0.87 $\rightarrow$ 0.84), while in the **MIXED** case the degradation varied from 0.89 $\rightarrow$ 0.82. This behavior is mainly due to the noisy labels produced by the ASR system (see [32]) i.e., 13% WER on *SOL-Cnt test set*.

Table 5: *Comparison of acoustic-based (aIB) and text-based diarization result on ATCO, PILOT, and MIXED subsets of SOL-Cnt test set.* †*acoustic aIB diarization system.* ††*proposed BERT model trained on* ***ALL_DATA*** *and evaluated on ground truth annotations (_GT) or ASR transcripts (_ASR).*

	DER (%)	JER (%)
Model	ATCO / PILOT / MIXED	ATCO / PILOT / MIXED
Acoustic_aIB†	14.8 / 13.9 / 13.1	15.6 / 13.5 / 25.5
BERT_GT††	**2.4 / 2.4 / 8.9**	**1.0 / 2.2 / 15.0**
BERT_ASR††	3.0 / 3.7 / 9.5	1.5 / 3.2 / 15.1

Breaking the paradigm, acoustic or text-based diarization? On challenging tasks such as ATC, where the rate of speech is high and contains mainly close-talk recordings, the standard acoustic-based diarization systems are prone to fail and merge two or more segments. An example is *SOL-Cnt* data set (see Table 2) where $\sim$38% of the test set contains more than one speaker or/and segment per utterance (i.e., 'Mixed'). We compare acoustic-based (aIB) and text-based diarization of *SOL-Cnt test set* in Table 5. As both systems use the same SAD, the DER using oracle SAD which is referred to as speaker error rate (SER) is reported. Here, we need to mention that acoustic-based diarization, does not evaluate SRD and evaluates SCD and clustering. For estimating the DER, we align the text with audio data and prepare the labeled segments from it. Using this alignment, the output of text-based diarization system is comparable to the acoustic-based diarization system. We found out that on noisy conditions, acoustic-based diarization mistakenly oversplit the segments with one speaker (ATCO/pilot). However, text-based diarization seems to be very robust on these segments (3.0/3.7% $\rightarrow$ 14.8/13.9% DER for ATCO/pilot). Even in the MIXED scenario, the BERT-based SRD and SCD systems (9.5% DER) extended with data augmentation outperform the acoustic-based model (13.1% DER) by 27% relative.

5. Conclusion

In this work, we demonstrated that acoustic-based tasks such as diarization can be enhanced or even replaced by natural language processing techniques, even when applied in challenging ATC scenarios. Our results, obtained on examples where SAD failed, validated this hypothesis, as presented in Table 4 and Table 5. Additionally, we developed a simple and flexible data augmentation pipeline for ATC task. To the authors' knowledge, this is the first time that a BERT-based system has been used for SCD of ATC data. This approach is capable of recognizing ATCO segments with $\sim$95% F1-score on SRD. The pilot segments were more challenging and the performance is $\sim$90% on SRD (Table 3). In the case of segments with more than one speaker (MIXED case), 89% F1-score was achieved on SCD. Text-based diarization shows a 27% relative improvement when compared to acoustic-based diarization. Future research should be directed at evaluating the fusion of acoustic and text-based

diarization and also relaxing the constraint that the number of speakers is fixed to two in current version.

6. References

[1] 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.

[2] M. Kleinert *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] Y. Lin, "Spoken instruction understanding in air traffic control: Challenge, technique, and application," *Aerospace*, vol. 8, no. 3, p. 65, 2021.

[4] 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 *Interspeech*, 2017.

[5] 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.

[6] M. Kocour, K. Veselỳ, 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.

[7] S. Madikeri and H. Bourlard, "Filterbank slope based features for speaker diarization," in *ICASSP*. IEEE, 2014, pp. 111–115.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[9] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge." in *Interspeech*, 2018, pp. 2808–2812.

[10] 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.

[11] N. Dawalatabad, S. Madikeri, C. C. Sekhar, and H. A. Murthy, "Incremental transfer learning in two-pass information bottleneck based speaker diarization system for meetings," in *ICASSP*. IEEE, 2019, pp. 6291–6295.

[12] T. J. Park and P. Georgiou, "Multimodal Speaker Segmentation and Diarization Using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks," in *Interspeech*, 2018, pp. 1373–1377.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, pp. 2493–2537, 2011.

[14] N. Mohapatra, N. Sarraf *et al.*, "Domain based chunking," *International Journal on Natural Language Computing (IJNLC) Vol*, vol. 10, 2021.

[15] J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, and R. Yangarber, "The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages," in *Proc. of the 6th Workshop on Balto-Slavic Natural Language Processing*, 2017, pp. 76–85.

[16] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *Proc. of the 27th International Conference on Computational Linguistics*, 2018, pp. 2145–2158.

[17] A. Sharma, S. Chakraborty, S. Kumar *et al.*, "Named entity recognition in natural language processing: A systematic review," in *Proceedings of Second Doctoral Symposium on Computational Intelligence*. Springer, 2022, pp. 817–828.

[18] A. Prasad, J. Zuluaga-Gomez *et al.*, "Grammar based identification of speaker role for improving atco and pilot asr." Idiap Research Institute, 2021, pp. 1–5.

[19] K. Ma, C. Xiao, and J. D. Choi, "Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks," in *Proceedings of ACL 2017, Student Research Workshop*. Association for Computational Linguistics, 2017, pp. 49–55.

[20] E. Han, C. Lee, and A. Stolcke, "Bw-eda-eend: Streaming end-to-end neural speaker diarization for a variable number of speakers," in *ICASSP*. IEEE, 2021, pp. 7193–7197.

[21] A. Khare, E. Han, Y. Yang, and A. Stolcke, "Asr-aware end-to-end neural diarization," in *ICASSP*. IEEE, 2022.

[22] O. Ohneiser, S. Sarfjoo, H. Helmke, S. Shetty, P. Motlicek, M. Kleinert, H. Ehr, and Š. Murauskas, "Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances," in *Interspeech*, 2021, pp. 3291–3295.

[23] B. Khonglah, S. Madikeri, S. Dey, H. Bourlard, P. Motlicek, and J. Billa, "Incremental semi-supervised learning for multi-genre speech recognition," in *ICASSP*, 2020.

[24] I. Nigmatulina, R. Braun, J. Zuluaga-Gomez, and P. Motlicek, "Improving callsign recognition with air-surveillance data in air-traffic communication," Idiap Research Institute. Idiap Research Institute, 2021, pp. 1–5.

[25] 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.

[26] S. S. Sarfjoo, S. Madikeri, and P. Motlicek, "Speech activity detection based on multilingual speech recognition system," in *Interspeech*, 2021, p. 4369–4373.

[27] 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.

[28] 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.

[29] Q. Lhoest *et al.*, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175–184.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[31] J. Zuluaga-Gomez, P. Motlicek, Q. Zhan, K. Veselỳ, and R. Braun, "Automatic Speech Recognition Benchmark for Air-Traffic Communications," in *Interspeech*, 2020, pp. 2297–2301.

[32] J. Zuluaga-Gomez, K. Veselỳ *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.

[33] J. Zuluaga-Gomez, I. Nigmatulina *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.

[34] N. Ryant *et al.*, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech 2019*, 2019, pp. 978–982.