BERTRAFFIC: BERT-BASED JOINT SPEAKER ROLE AND SPEAKER CHANGE DETECTION FOR AIR TRAFFIC CONTROL COMMUNICATIONS

Juan Zuluaga-Gomez [⋆,†,‡], *Seyyed Saeed Sarfjoo* [†], *Amrutha Prasad* [†,¶], *Iuliia Nigmatulina* [†], *Petr Motlicek* [†,¶], *Karel Ondrej* [¶], *Oliver Ohneiser* [§], *Hartmut Helmke* [§]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[¶] Brno University of Technology, Brno, Czech Republic
[§] German Aerospace Center (DLR), Institute of Flight Guidance, Braunschweig, Germany

2022 IEEE Spoken Language Technology Workshop (SLT) | 979-8-3503-9690-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SLT54892.2023.10022718

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. One common challenge is speech activity detection (SAD) and speaker diarization (SD). In the failure condition, two or more segments remain in the same recording, jeopardizing the overall performance. We propose a system that combines SAD and a BERT model to perform speaker change detection and speaker role detection (SRD) by chunking ASR transcripts, i.e., SD with a defined number of speakers together with SRD. The proposed model is evaluated on real-life public ATC databases. Our BERT SD model baseline reaches up to 10% and 20% token-based Jaccard error rate (JER) in public and private ATC databases. We also achieved relative improvements of 32% and 7.7% in JERs and SD error rate (DER), respectively, compared to VBx, a well-known SD system.[1]

Index Terms— Text-based speaker 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 (names of the aircraft). Considering that a big portion of this communication is transmitted via voice messages, previous studies proposed to apply automatic speech recognition (ASR) to automatically extract the corresponding transcripts and high-level entities. In recent years, the ASR systems were shown to reach maturity in reducing ATCO's workload, but for research-only scenarios. Examples are AcListant®-Strips [1] and MALORCA[2] projects. The later shows that novel data-driven machine learning approaches enable costly adaptations to different airport environments [2]. Lin [3, 4] reviews ten tasks on spoken instruction understanding of air traffic control (ATC) data. Semi-supervised learning has also been explored on the framework of ATC [5]. HAAWAII[3] and SOL-Cnt projects focus on developing a reliable and adaptable ASR engine for transcribing ATCO-pilot ATC communications. Previous work has concluded that higher accent variability and noise level cause ASR systems to yield up to two times higher word error rates (WER) for pilots' utterances compared to ATCOs' utterances [6]. In addition, close and cross-talk between ATCO and pilots induce acoustic-based speaker diarization (SD) systems to yield non-acceptable performances. All this together jeopardizes the speaker change detection (SCD) step and subsequently the ASR system ends 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, words in bold). [†]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

Already existent acoustic-based SD systems, like [7] or end-to-end neural-based SD [8], show promising performances for many applications. However, in ATC communications, given its limitations such as high speaker rate, close-talk, and noise levels, relying solely on the acoustic level has shown to be insufficient. Additionally, standard SD systems add one layer of complexity to the whole ATC

The work was supported by SESAR Joint Undertaking under Grant Agreement No. 884287 - HAAWAII (**H**ighly **A**utomated **A**ir traffic controller **W**orkstations with **A**rtificial **I**ntelligence **I**ntegration). The work was also partially supported by CleanSky Joint Undertaking under Grant Agreement No. 864702 - ATCO2 (Automatic collection and processing of voice data from air-traffic communications).

⋆Corresponding author: juan-pablo.zuluaga@idiap.ch

[1]Our code is stored in the following public GitHub repository: `https://github.com/idiap/bert-text-diarization-atc`

[2]`https://www.malorca-project.de`

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

pipeline,[4] weakening the flexibility to transfer the already tuned pipelines to other environments (e.g., noise level variation or new speakers' accents). That is why applying SD solely on the text level stands as an interesting solution to target these disadvantages. Additionally, the proposed SD system is speaker-agnostic because it fed with text data. This, can drastically reduce the chance of speaker identification as we remove the possibility to obtain the speaker identity from acoustic data or features.

1.2. Contribution

In this work, we fine-tune a pre-trained BERT model [9] to jointly perform tagging and chunking for SCD and speaker role detection (SRD). Chunking allows splitting sentences into tokens (or words) and then merging them in meaningful subgroups. In our case, a phrase (or entity) is composed of a full single-speaker utterance, where either ATCO or pilot is the 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). In short, our approach simplifies the standard SD pipeline, moving up the task from the acoustic level to text level, i.e., post ASR. We stack the BERT model on top of a speech activity detection (SAD) module to create a text-based SD, which from now on we call ***'BERT SD system'***. Our approach is proved on public and private databases. We developed a simple yet effective data augmentation pipeline to counteract the class imbalance within the train sets. The BERT SD system (i.e, combination of SAD, SRD, and SCD) yielded acceptable token-based JER of about 10% for seen domains (i.e., text transcripts provided to fine-tune the BERT model) and no more than 20% JER for databases that have not seen during training (in this case, the model has been fine-tuned on the SD task with other in-domain text data). Finally, we also experimented by directly feeding the BERT-based SD with transcripts generated by our in-domain hybrid-based ASR system for ATC [10, 11]. We obtained competitive performances compared to acoustic-based SD baselines.

2. RELATED WORK

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

Traditional acoustic-based diarization: feature representations of speakers are one of the main factors in the accuracy of a SD system. Mel frequency cepstral coefficients (MFCCs) are commonly used for the task of SD. 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 [12]. The agglomerative information bottleneck (aIB) approach to SD has shown competitive performance [13]. 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 [14]. For finding the speaker clusters in a sequence of x-vectors, the variational Bayesian hidden Markov model (VBx) was investigated in [15, 7]. For continuously learning speaker discriminative information, "Remember-Learn-Transfer" was proposed in [16]. Applying lexical and acoustic information for SD was investigated in [17].

Neural-based diarization: in the last years, there has been an increasing interest in end-to-end (E2E) and sequence-to-sequence architectures for different speech-related tasks. SD and its derivates, e.g., SCD, have also seen the benefits from this trend. For example, [18] builds on top of their proposed baseline for SD (ASR and SD are run in parallel and then the output is merged). They perform joint ASR and SD, claiming that word-level DER can be improved up to 15.8% in cross-domain evaluations. Afterward, the end-to-end neural diarization (EEND) was introduced in [8], where a full SD model is trained jointly to perform extraction and clustering. Later, the same authors upgraded the system by replacing the bidirectional long short-term memory (BLSTM) layers by self-attention modules [19]. Subsequent work has targeted EEND for unknown number of speakers [20], SD for long conversations [21], streaming EEND [22], SD constrained by turn detection (i.e., SCD) in [23], or even leveraging EEND for ASR [24].

Text-based speaker role detection: early text-based techniques for SRD or SCD relied on handcrafted lexicons, dictionaries, and rules. They are prone to human errors and not robust against noisy labels, e.g., produced by standard ASR systems (e.g., [25]). Collobert et al. [26] introduced machine learning methods for text processing in part-of-speech tagging, chunking, and semantic role labeling. In [27], domain-based chunking of sentences is addressed, which is similar to the approach proposed in this paper. In general, chunking is used to parse phrases from unstructured text. In our case, tagging and chunking an ATC utterance allows us to perform jointly SCD and SRD. 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 [28, 29, 30]. Examples of named entities in ATC communications are *callsigns*, *command types*, etc. These entities carry rich information that gives cues about the speaker's role (ATCO or pilot). A recent work aligned to ATC domain is reviewed in [31]. Here, a grammar-based approach performs SRD on single-speaker utterances. In [32] a text-based SRD for multiparty dialogues is proposed, but limited to SRD. Finally, text-based diarization has been proposed in the past by [22, 24]. However, these previous works do not take into account the text structure, grammar, and syntax.

Contrasting with previous work: different to other systems, e.g., EEND or traditional acoustic-based SD, our model is fed directly with text data (for instance, transcripts). The field of ATC holds some limitations and advantages regarding SD, where already existent acoustic-based EEND systems could fail. Some limitations are: ATC audio is noisy (below 15 dB SNR) with close and cross-talk speech. Some advantages are: the number of speaker roles are known (in our case two, ATCO and pilot) and the grammar between the two speaker roles slightly differs. Our main idea is to leverage those advantages in order to show that a fully text-based joint SCD and SRD system can perform on par or even better than traditional acoustic-based SD. As a clarification, similar scenarios where our approach can be applied are call-centers or patient–physician conversations, where the number of speaker roles are defined beforehand and their grammar structure also differs.

To summarize, the main difference between our BERT SD system and EEND roots on the fact that we use a standard BERT model [9] fine-tuned to ATC text data instead of crafting a SD neural network system. BERT[5] is known for its ease and powerful ability to be fine-tuned on many tasks and corpus with minimum effort

[4] A standard ATC pipeline is composed of signal processing, SAD and SD, ASR, natural language understanding and post-processing.

[5] We use *BERT-base-uncased* (110M params) for all the experiments.

Table 2. Amount of train and test data (# train utterances / # test utterances) for each class. *ATCO* and *pilot* columns are single-speaker samples, while *Mixed* are utterances with two or more segments. †real-life ATC set where speech activity detection failed.

Database	ATCO	Pilot	Mixed	Ref
Private databases				
SOL-Cnt†	662 / 138	945 / 204	535 / 205	[33]
HAAWAII	18724 / 1954	21099 / 2299	- / -	[34]
Public databases				
ATCO2	- / 1772	- / 1350	- / -	[11]
LDC-ATCC	12694 / 1515	14216 / 1446	- / -	[35]
UWB-ATCC†	4577 / 1157	6669 / 1713	735/174	[36]

(e.g., amount of labeled data). It also performs well in low-resource scenario, which is the case in ATC. Finally, as our system removes the 'acoustic level' complexity and moves it to the text level, we demonstrate that mapping to the target domain when we have specific speaker roles is more efficient in the text level. For instance, data augmentation on text is simpler than on the acoustic level or one can modify easily the training data to adapt it to another scenario by merely altering the text.

3. DATABASES AND EXPERIMENTAL SETUP

This research experiments on five databases in the English language with various accents and data quality. With the aim of encouraging open research on ATC (which has lagged behind due to privacy clauses and contracts[6]) we identified and experimented with three public databases, as referenced in Table 2. All experiments use 10% of the train set as validation set. We release a GitHub repository [7] with training scripts to replicate the results on the public databases. To the author's knowledge, this is the first open release of code in the field of natural language processing for air traffic control.

3.1. Private databases

SOL-Cnt: private database recorded and collected over EU-funded industrial research project that 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[8] [33].

HAAWAII: private data set that has been collected and annotated from ATC communications from London and Icelandic airspace.[9] This data is of higher quality ($\geq$ 15 dB SNR) compared to SOL-Cnt. All the data is correctly split, i.e., one speaker per segment. Previous benchmark and results are covered in [37, 34].

3.2. Public databases

LDC-ATCC: public ATC corpus gathered from three different airports in the US. **LDC-ATCC**[10] comprises recorded speech with the aim of supporting research in robust ASR. The recordings contain several speakers, and they were collected over noisy channels. The database is formatted in NIST Sphere format, where full transcripts, start and end times of each transmission are provided [35].

UWB-ATCC: public ATC corpus containing recordings of communication between ATCOs and pilots. The speech is manually transcribed and labeled with the information about the speaker (pilot/controller, not the full identity of the person). The audio data is single channel sampled at 8kHz. Similar to SOL-Cnt, UWB-ATCC contains around 900 utterances where segmentation failed and two or more segments and/or speakers ended up in the same recording. This database can be downloaded for free in their website [11] [36].

ATCO2 corpus: public ATC corpus recently released in ATCO2 project.[12] ATCO2 developed a pipeline to pseudo-annotate (ASR transcripts, language and diarization labels) large amounts of ATC speech audio for training robust ASR models. We use the entire database only as test set (over 4000 utterances), thus we consider this as an out-of-domain evaluation. The ATCO2 corpus is one of the few open-source and public databases that has been used by other researchers to benchmark their ASR engines [11, 38]. The full corpus is available for purchase through ELDA in `http://www.elra.info/en/catalogues/`.

3.3. Annotation protocol

In addition to manual speech transcripts, speaker labels and time segmentation (e.g., ATCO/pilot/mixed) are also available. The BERT model starts by tagging each word of the transcript (ground truth or ASR transcript) with a set of tags that follows the well-known *IOB format* (Inside-Outside-Beginning). In IOB format, each entity (a full sentence in our case) is composed of two tags: (i) the *Beginning* defines which token/word is the start of the sentence ***'B-'***, and (ii) the *Inside* tag ***'I-'*** defines which tokens/words belongs to that specific sentence. We define ATCO recordings as *Speaker1*, while pilot segments as *Speaker2* (green and red in Figure 1, respectively). We do not use the *Outside* tag, because we know that each word is always from one of two predefined speakers. In total, we have four tags, two per class (ATCO and pilot).

3.4. Data augmentation

We implemented a simple yet effective data augmentation pipeline to counteract the class imbalance in the train sets (see Table 2). First, we split the training sets on either *ATCO* (speaker 1) or pilot (speaker 2) subset. Then, we generate new sentences from the initial set of utterances for each database (e.g., HAAWAII $\sim$39k utterances). Each new sample depends on: (i) the number of sentences to be concatenated, and (ii) the speaker label of each sentence. In general, a new sample is composed of one to four sentences, each with an equal chance of being drawn from the ATCO or pilot dictionary. The process is repeated until gathering $\sim$350 MB of text data ($\sim$1M sentences). We emphasize that in ATC, there is no need to have a correlation between previous and next sentences/utterances. This is due to the fact that speaker 1 (ATCO) communicates to several speakers 2 (pilots). The stream of information received and transmitted by the speakers is not dependent on 'left' or 'right' context. Therefore, concatenating various segments randomly would not degrade

[6] Nearly all ongoing and former projects in the area of ATC prohibit the release of code and models due to privacy issues.

[7] `https://github.com/idiap/bert-text-diarization-atc`

[8] PJ.16-W1-04: `https://www.sesarju.eu/projects/cwphmi`

[9] `https://www.haawaii.de`

[10] `https://catalog.ldc.upenn.edu/LDC94S14A`

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

[12] ATCO2 corpus: `https://www.atco2.org/data`

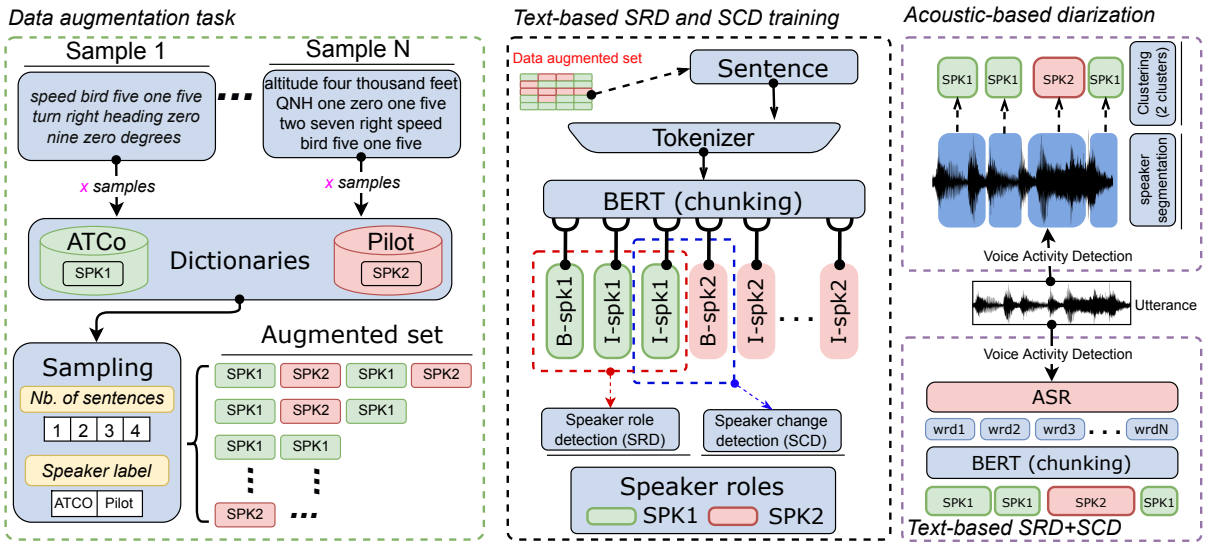

Fig. 1. **Left block:** proposed data augmentation pipeline. Augmented samples contain between one and four utterances (probabilities of 40%, 30%, 20% and 10% for one to four, respectively). New sentences have equal chance to be sampled from the ATCO or pilot dictionary. **Central block:** proposed pipeline to fine-tune a BERT model that performs tagging and chunking for joint SCD and SRD. **Right block:** proposed approach to compare acoustic-based SD (VBx) and BERT joint SRD and SCD.

substantially the WERs.[13] The left block in Figure 1 depicts the proposed data augmentation pipeline.

3.5. Modules

The performance of our BERT-based SRD and SCD system is contrasted with a standard acoustic-based SD system. We use an out-of-the-box VBx system to evaluate the *SOL-Cnt* and *UWB-ATCC* test sets, which contain real-life ATC audio where segmentation failed. For both, BERT and acoustic-based SD systems, we use the same multilingual ASR-based SAD module [39] 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 pre-trained BERT model [9] downloaded from HuggingFace [40, 41]. The model is later fine-tuned with either the original or the augmented databases, on the tagging and chunking task (following *IOB* format). We append a linear layer with a dimension of 4 (following the classes structure from Section 3.3) on top of the last layer of the BERT model. Then, we fine-tune each model on an NVIDIA GeForce RTX 3090 for 3k steps, with a learning rate scheduler that first warms up the learning rate until $\gamma = 5\text{e}{-5}$ for 500 steps, and then it linearly decays. We employ AdamW [42] optimizer (β_1=0.9, β_2=0.999, ϵ=1e$-$8) and dropout [43] of $dp = 0.1$ for the attention and hidden layers. We use GELU activation function [44]. We train all models with batch size of 32, and gradient accumulation of 2, i.e., effective batch size of 64.

Acoustic-based diarization: for details of the VBx model, the reader is referred to [7]. This model uses a Bayesian hidden Markov model (BHMM) to find speaker clusters in a sequence of x-vectors. Here, the x-vector extractor uses DNN architecture based on ResNet101. The input to the ResNet is 64 log Mel filter bank features extracted every 10 msec using 25 msec window. In the first step, Agglomerative Hierarchical Clustering (AHC) is applied to the extracted x-vectors. Then, Variational Bayes HMM over x-vectors is applied using the AHC output. For achieving the best performance on the database with short duration files with a maximum of two speakers, we tuned the probability of not switching speakers between frames (loopP) and speaker regularization coefficient (Fb) to 0.7 and 6, respectively.

Automatic speech recognition: a state-of-art hybrid-based ASR system for ATC speech was developed with Kaldi toolkit [45]. 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 audio from both speakers (i.e., ATCO and pilot). The training recipe and databases (including the train sets in Table 2) are covered in [11, 46, 47, 37].

3.6. Evaluation protocol

The experiments are prepared to answer three questions: (i) how reliable is the BERT-based SRD and SCD system on ground truth transcripts? (ii) How is the performance impacted when using automatically generated (ASR) transcripts instead of ground truth transcripts?[14] And, (iii) which system performs better on real-life ATC speech data, text or acoustic-based SD?

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 which is defined in Equation 1:

$$DER = \frac{\text{false alarm} + \text{miss detection} + \text{speaker confusion}}{\text{Total duration of speech in the reference file}}, \quad (1)$$

where false alarm is the duration of non-speech incorrectly classified as speech, missed detection is the duration of speech incorrectly

[13]We measure WERs by decoding with the in-domain ASR system the original and augmented test sets to corroborate this assumption.The relative WER degradation was less than 1% for all test sets.

[14]This is a real-life scenario where ASR transcripts are fed to the BERT SD system instead of ground truth transcripts.

Table 3. Jaccard error rate (JER) in percentages (%) from predictions using different train (column 1) and test sets. All the experiments use the same model (BERT-base-uncased) and same set of hyperparameters. We report the mean of five runs with different seeds and its standard deviations (mean ± STD). **Bold** refers to the best performance over public databases, while underline denotes the highest performance per column. Metrics reported on token level of ground truth transcripts.

Model		Public			Private	
Database	**# samples**	**ATCO2**	**UWB-ATCC**	**LDC-ATCC**	**HAAWAII**	**SOL-Cnt**
Public databases						
LDC-ATCC	26.9k	31.3 ± 2.4	35.8 ± 2.0	8.1 ± 0.7	28.7 ± 3.1	52.6 ± 1.3
UWB-ATCC	11.2k	21.6 ± 0.7	$\underline{\mathbf{10.7 \pm 0.6}}$	18.7 ± 2.6	15.2 ± 1.4	$\underline{\mathbf{18.7 \pm 1.7}}$
↪ + LDC-ATCC	38.1k	$\mathbf{19.8 \pm 0.9}$	11.3 ± 0.4	$\underline{\mathbf{7.1 \pm 1.3}}$	$\mathbf{14.2 \pm 1.4}$	24.0 ± 1.9
Private database						
HAAWAII	39.8k	23.9 ± 0.6	22.3 ± 1.7	14.1 ± 1.2	6.5 ± 0.7	48.5 ± 1.4
↪ +LDC+UWB	77.9k	$\underline{17.5 \pm 0.2}$	11.5 ± 0.5	7.5 ± 0.6	$\underline{6.2 \pm 0.3}$	26.8 ± 2.0

classified as non-speech, confusion is the duration of speaker confusion, and total is the total duration of speech in the reference. JER is a recently proposed metric [48] that avoids the bias towards the dominant speaker, i.e., evaluating equally all speakers. The JER is defined in Equation 2:

$$JER = 1 - \frac{1}{\#\text{speakers}} \sum_{\text{speaker}} \max_{\text{cluster}} \frac{|\text{speaker} \cap \text{cluster}|}{|\text{speaker} \cup \text{cluster}|}, \quad (2)$$

where speaker is the selected speaker from reference and $max_{cluster}$ is the cluster from the system with maximum overlap duration with the currently selected speaker.

Speaker role detection: we evaluate SRD with JER on the token level (which is more aligned to SD) on the five proposed test sets. To clarify, **SOL-Cnt** and **UWB-ATCC** databases contain utterances with more than one speaker per utterance. Thus, we test the SD capabilities of the proposed BERT-based system on these two test sets. Results are shortlisted in Table 4. We first analyzed the model performance on the ideal case, i.e., we used the ground truth audio annotations to obtain JERs per test set, thus assuming we have access to a perfect ASR system (0% WER). These results are listed in Table 3. We employ the Scikit-learn[15] Python library to calculate these scores.

Speaker change detection: in addition to SRD, the BERT system performs SCD, i.e., central block in Figure 1. We evaluated this task with DER and JER on one private (**SOL-Cnt**) and one public (**UWB-ATCC**) test set, which contains utterances with one or two speakers. The *MIXED* column in Table 4 list the results corresponding to SCD only on the multi-speaker segments. For creating the segments from the BERT-based SCD system, we used forced alignment between audio and ground truth text using the trained ASR module. This module is explained in Section 3.5. Similarly, time information from the ASR output transcripts was used to create the segments of the BERT-based SD system on the ASR transcripts.

4. RESULTS AND DISCUSSION

Baseline performance of BERT SD: we discuss the results listed in Table 3. Here, we aim at evaluating two aspects of the BERT SD system. First, we assess how well the model behaves on out-of-domain corpora. We fine-tune BERT models on each database and evaluate it on all five test sets. We call this: *transferability* between corpora. Second, we establish baselines on both, public and private databases. Each model is fine-tuned five times with different seeds, hence we report the mean and standard deviation across runs. Not to our surprise, test data that matched the fine-tuning one performed particularly well. LDC-ATCC and UWB-ATCC test sets reached less than 10% JER, while ∼20% JER for ATCO2.

One aspect that can shed light on new research is how public databases transfer to private ones. This can help future research to set a starting point, thus reducing the costs inherit by developing tools from scratch, e.g., SD system for ATC. We noted that UWB-ATCC corpus was more challenging for the BERT SD model compared to LDC-ATCC and HAAWAII corpora (6.5% and 8.1% JER, respectively). However, this system performed consistently better on all the other test sets, if we compare the model fine-tuned on UWB-ATCC versus the ones on LDC-ATCC and HAAWAII. We believe that the transferability to new domain of UWB-ATCC corpus is higher compared to LDC-ATCC and HAAWAII (see 'UWB-ATCC' row in Table 3 and compare it with LDC-ATCC or HAAWAII).

Does adding more data help? Here, we evaluate the BERT SD system by performing an ablation where the amount of fine-tuning data is incremental. In total, 9 models per database are studied, as depicted in Figure 2 (each data point represents one model). We report token-based JERs which are more aligned to standard SD. For the public databases, we obtained 65, 43, and 37% relative improvement in JERs on LDC, UWB, and ATCO2, respectively, by scaling up the fine-tuning data from 100 to 2000 samples. This number goes up to more than 50% relative JERs improvement if we use 10k samples (69% relative improvement for LDC). We note the same behavior on the private databases. At least 50% relative improvement is seen by scaling up the data from 100 samples → 2000 samples, on both, HAAWAII and SOL-Cnt experiments. To our surprise, UWB-ATCC models transfer particularly well on the two out-of-domain test sets (i.e., ATCO2 and SOL-Cnt). This gives insights that our approach works well on both, public and private databases. We believe this is an acceptable starting point for the future research on text-based SRD and SCD (not only aligned to ATC).

Robustness of BERT speaker diarization on ASR transcripts: we evaluated the BERT SD system on *SOL-Cnt* and *UWB-ATCC* test sets, which contain utterances with more than one speaker (*mixed*

[15]We use weighted Jaccard error rate score. It calculates metrics for each class (i.e., ATCO and pilot), and finds their average weighted by support (the number of true instances for each class). This accounts for label imbalance.

Table 4. Comparison of acoustic-based VBx SD and text-based SD on *ATCO*, *PILOT*, and *MIXED* subsets of SOL-Cnt, and UWB-ATCC test sets. **Bold** refers to the best performance. †the performance of acoustic diarization system. ††proposed BERT model trained on all the available data with data augmentation and evaluated on ground truth annotations (_GT) or ASR transcripts (_ASR).

	Sol-Cnt test set		**UWB-ATCC test set**	
	DER (%) ↓	**JER (%) ↓**	**DER (%) ↓**	**JER (%) ↓**
Model	AT / PI / MIX	AT / PI / MIX	AT / PI / MIX	AT / PI / MIX
Acoustic-based speaker diarization				
Acoustic_VBx†	5.8 / 7.8 / 10.3	7.0 / 10.9 / 22.2	**0.8** / **1.2** / 14.4	**0.6** / **0.7** / 39.4
Text-based speaker diarization				
BERT_GT††	**2.4** / **2.4** / **8.9**	**1.0** / **2.2** / **15.0**	1.2 / 1.7 / **5.8**	1.1 / 1.1 / **16.6**
BERT_ASR††	3.0 / 3.7 / 9.5	1.5 / 3.2 / 15.1	1.6 / 1.5 / 6.9	1.2 / 1.2 / 20.1

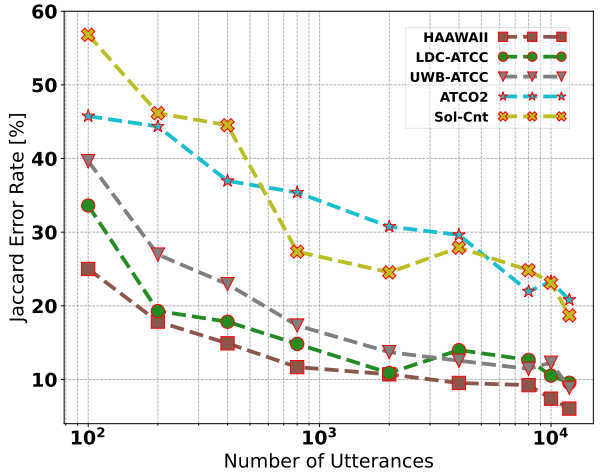

Fig. 2. Jaccard error rates (JER) in percentages (%) for nine models fine-tuned with increased amount of samples per database. We evaluate models on two configuration. HAAWAII, LDC-ATCC and UWB-ATCC are *in domain experiments*, which means that the train and test splits are from the same database. ATCO2 and SOL-Cnt are *out of domain test sets,* i.e., the train and test data differs. For the two later (blue and yellow dashed lines), we report the results of the model fine-tuned with UWB-ATCC database.

subset). The BERT SD system is fed with the 1-best transcript obtained from our in-domain hybrid-based ASR system [37]. Table 4 highlights the main results for the BERT SD model, an additional line for *'ASR output'*. In the single-speaker case (either ATCO or pilot), the degradation (ASR transcripts instead of ground truth text) in SD from the BERT SD was no more than 1% absolute JER and DER (worse, Pilot subset 2.4 → 3.7% DER reduction in *SOL-Cnt* set). In the **MIXED** case, the degradation varied 0.1% JER and 0.6% DER absolute in *SOL-Cnt* set, and 3.5% JER and 1.1% DER absolute in *UWB-ATCC* set. This behavior is mainly due to the noisy labels produced by the ASR system (see [38]), i.e., 13% and 14% WER on *SOL-Cnt* and *UWB-ATCC* test sets.

Breaking the paradigm, acoustic or text-based speaker diarization? On challenging tasks such as ATC, where the rate of speech is high and contains mainly close-talk recordings, the standard acoustic-based SD systems are prone to fail and merge two or more segments together. An example is *SOL-Cnt* database (see Table 2) where ∼38% of the test set contains more than one speaker or/and segment per utterance (i.e., *'Mixed'*). We compare acoustic-based and BERT SD on private (*SOL-Cnt*) and public (*UWB-ATCC*) test sets. Similar to *SOL-Cnt*, *UWB-ATCC* set contains more than one speaker per utterance. We list the results in Table 4. In order to contrast both approaches, we compute the JER on the extracted segments, not on the text-level tokens (as done before). Both systems use the same SAD for segmentation. The acoustic-based SD, uses the Hungarian algorithm [49] for assigning the system clusters to the reference speakers. As a result, it evaluates SCD and clustering without identifying the speaker roles. For estimating the DER, we align the text with audio data and prepare the labeled segments from it. Using this alignment, the output of the BERT SD system is comparable to the acoustic-based diarization system. For computing the scores in all systems, the collar of 150 msec was considered. We found out that in noisy conditions, acoustic-based SD mistakenly oversplits the segments with one speaker (either ATCO or pilot). However, the BERT SD seems to be very robust on these segments (3.0/3.7% → 5.8/7.8% DER for ATCO/pilot of *SOL-Cnt* test set). Even in the mixed scenario of this set, the BERT SD system (9.5% DER) extended with data augmentation outperformed the acoustic-based model (10.3% DER) by 7.7%, relatively. On a cleaner set with shorter segments, VBx system shows the best performance on the segments with one speaker. However, in the mixed segments, the BERT SD system outperformed the VBx by a marginal improvement.

5. CONCLUSION

In this work, we demonstrated that acoustic-based tasks such as speaker diarization can be enhanced or even replaced by natural language processing techniques. Even including challenging tasks such as SD for ATC communications. Our results, obtained on examples where SAD failed, validated this hypothesis, as presented in Table 3 and Table 4. Additionally, we developed a simple and flexible data augmentation pipeline for ATC text data. To the authors' knowledge, this is the first time that a BERT-based SD could fully replace an acoustic-based SD in the field of ATC. We evaluated our approach on public and private datasets in the ATC domain. Our BERT SD model reached up to 10% and 20% token-based JER in public and private ATC databases. We compared our model with the well-known acoustic-based SD system (VBx). On the noisy sets, VBx oversplits the segments with one speaker, however, the BERT SD system shows robust performance on these segments. In addition, BERT SD model outperforms VBx by a large margin in segments with more than one speaker (*MIXED*). Finally, we also performed an ablation of the amount of data samples versus performance.

6. REFERENCES

[1] Hartmut Helmke, Oliver Ohneiser, Jörg Buxbaum, and Chr 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] Matthias 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] Yi Lin, "Spoken instruction understanding in air traffic control: Challenge, technique, and application," *Aerospace*, vol. 8, no. 3, pp. 65, 2021.

[4] Lin Yi, Ruan Min, Cai Kunjie, Li Dan, Zeng Ziqiang, Li Fan, and Yang Bo, "Identifying and managing risks of ai-driven operations: A case study of automatic speech recognition for improving air traffic safety," *Chinese Journal of Aeronautics*, 2022.

[5] Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan, Gyorgy Szaszak, Youssef Oualil, and Hartmut Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Interspeech*, 2017.

[6] Thomas Pellegrini, Jérôme Farinas, Estelle Delpech, and François Lancelot, "The airbus air traffic control speech recognition 2018 challenge: towards atc automatic transcription and call sign detection," *arXiv preprint arXiv:1810.12614*, 2018.

[7] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget, "Bayesian HMM clustering of x-vector seq uences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.

[8] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

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

[10] Juan Zuluaga-Gomez, Iuliia 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.

[11] Martin Kocour, Karel Veselỳ, Igor Szöke, Santosh Kesiraju, Juan Zuluaga-Gomez, Alexander Blatt, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlíček, Dietrich Klakow, et al., "Automatic processing pipeline for collecting and annotating air-traffic voice communication data," *Engineering Proceedings*, vol. 13, no. 1, pp. 8, 2021.

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

[13] Deepu Vijayasenan, Fabio Valente, and Hervé 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.

[14] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji 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.

[15] Fabio Valente, Petr Motlicek, and Deepu Vijayasenan, "Variational bayesian speaker diarization of meeting recordings," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4954–4957.

[16] Nauman Dawalatabad, Srikanth Madikeri, C Chandra Sekhar, and Hema A Murthy, "Incremental transfer learning in two-pass information bottleneck based speaker diarization system for meetings," in *ICASSP*. IEEE, 2019, pp. 6291–6295.

[17] Tae Jin Park and Panayiotis Georgiou, "Multimodal Speaker Segmentation and Diarization Using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks," in *Interspeech*, 2018, pp. 1373–1377.

[18] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, "Joint speech recognition and speaker diarization via sequence transduction," *Proc. Interspeech 2019*, pp. 396–400, 2019.

[19] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[20] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *arXiv preprint arXiv:2005.09921*, 2020.

[21] Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison W Cottrell, "Speech recognition and multi-speaker diarization of long conversations," 2020.

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

[23] Wei Xia, Han Lu, Quan Wang, Anshuman Tripathi, Yiling Huang, Ignacio Lopez Moreno, and Hasim Sak, "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8077–8081.

[24] Aparna Khare, Eunjung Han, Yuguang Yang, and Andreas Stolcke, "Asr-aware end-to-end neural diarization," in *ICASSP*. IEEE, 2022.

[25] Fabio Valente, Deepu Vijayasenan, and Petr Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4416–4419.

[26] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, pp. 2493–2537, 2011.

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

[28] Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman 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.

[29] Vikas Yadav and Steven 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.

[30] Abhishek Sharma, Sudeshna Chakraborty, Shivam 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.

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

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

[33] Oliver Ohneiser, Saeed Sarfjoo, Hartmut Helmke, Shruthi Shetty, Petr Motlicek, Matthias Kleinert, Heiko Ehr, and Šarūnas Murauskas, "Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances," in *Interspeech*, 2021, pp. 3291–3295.

[34] Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan, "How does pre-trained wav2vec2.0 perform on domain shifted asr? an extensive benchmark on air traffic control communications," *IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar*, 2023.

[35] John Godfrey, "The Air Traffic Control Corpus (ATC0) - LDC94S14A," 1994.

[36] Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindřich Matoušek, Jan Romportl, and Pavel Ircing, "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 449–464, 2019.

[37] Iuliia Nigmatulina, Juan Zuluaga-Gomez, Amrutha Prasad, Seyyed Saeed Sarfjoo, and Petr Motlicek, "A two-step approach to leverage contextual data: speech recognition in air-traffic communications," in *ICASSP*, 2022.

[38] Juan Zuluaga-Gomez, Karel Veselỳ, et al., "Automatic call sign detection: Matching air surveillance data with air traffic spoken communications," in *Multidisciplinary Digital Publishing Institute Proceedings*, 2020, vol. 59.

[39] Seyyed Saeed Sarfjoo, Srikanth Madikeri, and Petr Motlicek, "Speech activity detection based on multilingual speech recognition system," in *Interspeech*, 2021, p. 4369–4373.

[40] Thomas 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*. 2020, pp. 38–45, Association for Computational Linguistics.

[41] Quentin 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.

[42] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

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

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

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

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

[49] Roy Jonker and Ton Volgenant, "Improving the hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.