



# Detecting English Speech in the Air Traffic Control Voice Communication

Igor Szöke, Santosh Kesiraju, Ondřej Novotný, Martin Kocour, Karel Veselý, Jan “Honza” Černocký

Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

szoke@fit.vutbr.cz

## Abstract

Developing in-cockpit voice enabled applications require a real-world dataset with labels and annotations. We launched a community platform for collecting the Air-Traffic Control (ATC) speech, world-wide in the ATCO<sup>2</sup> project. Filtering out non-English speech is one of the main components in the data processing pipeline. The proposed English Language Detection (ELD) system is based on the embeddings from Bayesian sub-space multinomial model. It is trained on the word confusion network from an ASR system. It is robust, easy to train, and light weighted. We achieved 0.0439 equal-error-rate (EER), a 50% relative reduction as compared to the state-of-the-art acoustic ELD system based on x-vectors, in the in-domain scenario. Further, we achieved an EER of 0.1352, a 33% relative reduction as compared to the acoustic ELD, in the unseen language (out-of-domain) condition. We plan to publish the evaluation dataset from the ATCO<sup>2</sup> project.

**Index Terms:** speech recognition, language detection, x-vector extractor, acoustic model, air-traffic communication, data collection, text embeddings, Bayesian methods

## 1. Introduction

ATCO<sup>2</sup> project<sup>1</sup> aims at developing a unique platform allowing to collect, organize and pre-process Air-Traffic Control voice communication – ATC data – recorded through the VHF channel<sup>2</sup>. ATC voice data is tied with a metadata collected through ADS-B channel (altitude, call-sign, GPS coordinates, etc.) and automatically processed. The goal is to create a labeled speech dataset that grows in time, and allows human annotators to correct automatic transcriptions and tag errors. The ATCO<sup>2</sup> dataset will be useful in research and development of in-cockpit voice enabled applications.

The ATCO<sup>2</sup> project primarily aims at delivering good quality, low Signal-to-Noise Ratio (SNR) English speech data. Collecting such data is challenging because the data feeders<sup>3</sup> are spread globally, and the data contains many local languages and dialects of English. As we do not have control over the recording conditions, the quality of the data may vary<sup>4</sup>

The ATC data is challenging for automatic speech processing [1], especially in our environment driven by volunteers:

- Various noise conditions. The majority of data is “clean” (82% of data has SNR above 10dB), but there are dia-

<sup>1</sup><http://atco2.org>

<sup>2</sup>Very High Frequency, Amplitude modulation in range of 108MHz - 137MHz

<sup>3</sup>Volunteers with antennas and SDR devices capturing the communication from the air and uploading to ATCO<sup>2</sup> servers.

<sup>4</sup>As we are launching the platform now (March 2021), we present our findings on data collected from Prague (LKPR) and Brno (LKTB) airports in Czechia – our bootstrap data feeders. On the other hand, we do not expect significant change in the data properties and quality worldwide.

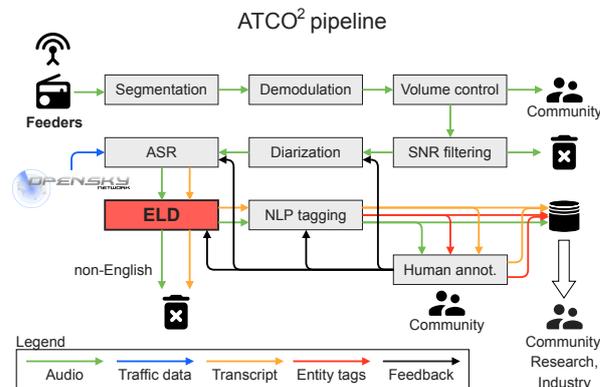


Figure 1: ATCO<sup>2</sup> data flow pipeline. The ELD block is used to filter our non-English speech. The automatically processed data are then annotated and checked by humans. Their feedback is used to adapt particular blocks using active learning approaches.

logues where the SNR difference between conversation sites is very large.

- Strongly accented English. The speakers’ English accent varies widely. From native speakers (pilots), through international accents (French, German, Russian, etc.) (pilots and ATCOs) to the strong local accent (Czech pilots and ATCOs in our initial data).
- Mixed words and phrases. The Czech ATCOs vocabulary is a mix of Czech and English words. They use standard greetings in Czech which can be a significant portion of an English sentence if the command is short. On the other hand, they use many English words (alphabet, some commands) in Czech sentences. Moreover they use a significant amount of “Czenglish” words. Our international partners report similar situation in their countries.

Because of the sparsity in VHF data (speech was present in 1% – 8% of time), and local language environment (English language was present in 70 – 80% of the segments), one VHF channel can generate about 20 – 30 hours of English speech per month.

### 1.1. Motivation

The ATCO<sup>2</sup> data processing pipeline should continuously process large amounts of data in an on-line manner (see Figure 1). The decision on (non)English speech should be light-weight and robust. It has to cope with the above mentioned channel, acoustic and lexical challenges while preserving an easy way to update a model to newer airports and countries.

We decided to explore the English language detection (ELD) based on an ASR output for the following two reasons: 1) there is about 20% of non-English speech, so run-

ning ASR on all the data does not bring a large overhead (compared to filtering non-English prior to the ASR block), 2) standard acoustic-based Language IDentification (LID) can be too data-consuming [2, 3, 4] and its accuracy suffers in bi-lingual code-switching environments [5]. The ASR-based ELD systems share similarities with phonotactic LID systems [6, 7, 8, 9]: the former relies on word confusion networks, the latter relies on phoneme sequences.

The most recent works targeting similar problems as ours are related to the AI-powered smart voice assistants. Here the motivation is to detect the language of short commands efficiently. The challenge increases in bilingual or multilingual environments with code switching. [10, 5] improved language detection in multilingual speech recognition for voice assistants: [10] experimented with combination of acoustic LID with ASR features such as 1-best, confidence, AM and LM scores. They were able to improve the language detection by 52% when adding ASR-based features to the acoustic LID. The motivation of [5] was to improve language detection in bilingual environment. They used combined acoustic and text (ASR 1-best) based features to detect language of a spoken command. They achieved about 56% relative error rate improvement using text-based LID compared to acoustic one and another 5% improvement with a combination of acoustic and text based features post-processed by a DNN.

The ATC domain is similar to voice assistants in the sense of noisy channel, bilingual environment (English and the local language) and code switching. On the other hand, we do not require a real time processing in our case.

This paper is aimed at the English Language Detection (ELD) in ATC domain. We describe an effective way of filtering out non-English language. We evaluate on our dataset described in Section 2. The description of acoustic ELD based on state-of-the-art x-vectors is given in Section 3. As the speech-to-text technology is one of our building blocks, we briefly discuss it in Section 4, we kindly ask the reader to follow [11] for more information. Description of various language detection systems based on the ASR output is presented in Section 5. The experiments and results are outlined in Section 6, and we conclude with directions for future work in Section 7.

## 2. Datasets

We found only one recent bi-lingual ATC dataset – English / Chinese ATCSpeech [12]<sup>5</sup>. Other datasets are English language only [13, 14, 15, 16, 17, 18] or multilingual but not in ATC domain [19].

Regarding the lack of multilingual ATC datasets with language tagging, we had to create one ourselves based on the collected ATCO<sup>2</sup> data. At the time of writing this paper, we had coverage of Czechia (Prague and Brno airports. i.e. English / Czech). To test robustness of our solution, we used HAAWAI project data from Iceland (Keflavík airport, i.e. English / Icelandic) as unseen (out-of-domain) data.

Assigning the language tag to particular segment (a short dialog between ATCO and pilot) was mainly based on the language of spoken numerals. We used the following rules:

- The default tag is English.
- Language of greetings is ignored.
- If numerals are spoken in Czech / Icelandic, the tag is Czech / Icelandic respectively.

<sup>5</sup>We have not found any suitable way how to acquire the corpus at the time of writing this paper.

- If there is “out of ATC grammar” speech in Czech / Icelandic, the tag is Czech / Icelandic.
- If no numerals are present and the commands / values are clearly in English, the tag is English.
- Otherwise, discard the speech segment and tag it as mixed (clearly both languages are present) or unknown language (cannot be determined).

The ATCO<sup>2</sup> ELD dataset is formed from two airports and comprises two subsets (see Table 1). Both subsets were recorded during a period of several months, and none of the recordings overlap in time.

Table 1: ATCO<sup>2</sup> ELD datasets. The first line is the Number of segments / Total length of speech [minutes]. LKPR – Czechia, Prague airport, LKTB – Czechia, Brno airport

	LKPR		LKTB	
	EN	CZ	EN	CZ
ENCZ-Dev	596/77	492/61	579/66	453/61
ENCZ-Eval	820/128	308/55	515/57	289/36

The HAAWAI ELD dataset is formed from a single airport and consists of two subsets (see Table 2). Both subsets were recorded during a period of two months and none of the recordings overlap in time.

Table 2: HAAWAI ELD datasets. The first line is the Number of segments / Total length of speech [minutes]. BIKF – Iceland, Keflavík Airport

	BIKF	
	EN	IS
ENIS-Dev	404/69	109/19
ENIS-Eval	414/80	111/24

We plan to enlarge the ATCO<sup>2</sup> database and add more languages during the project. We intend to publish the datasets during 2021. The HAAWAI dataset is proprietary and the project agreement does not allow to publish it.

## 3. Baseline Acoustic Based English Language Detection

Our baseline acoustic ELD system is based on neural network x-vector technology, initially introduced for speaker verification [20]. NN as an x-vector extractor was shown to be very robust in noisy environments [21, 22]. The original x-vector topology was successfully applied as a language recognizer [4].

Our system replaces the Time Delay Neural Network (TDNN) topology with a more efficient topology based on ResNet [23] with 18 layers and 64 filter bank channels as the acoustic features. Softmax was used as the output layer non-linearity with a class for each language.

Our training set contains data from NIST LRE 2009–2017 [24, 25, 26, 27], Babel [28], NIST SRE16 [29], and Fisher English [30]. The data was augmented using Kaldi [31] pipeline [32], without the babble noise branch. The training data comprises of 62 languages with 644k segments, totalling 21.6k hours in duration.

The x-vector embeddings are extracted and fed into Gaussian linear classifier (GLC) for English and non-English classification. Experimental details are given in Section 6.

## 4. Speech-to-text

The English ASR was trained on 7 corpora: AIRBUS [13], HIWIRE [33], LDC ATCC [14], MALORCA [18], N4 NATO [19], ATCOSIM [15], UWB ATCC [16, 17]. The total duration of audio was 136.3 hours.

We used hybrid speech-to-text recognizer based on Kaldi [31], trained with Lattice-free MMI training [34]. The neural network had 6 convolutional (`conv-relu-batchnorm-layer`) layers, followed by 9 semi-orthogonal components (`tdnnf-layer`) [35], and the model has two pre-final and output layers. In total, the model has 12.93 million trainable parameters, and the number of left bi-phone tied-states is 1680. The input features are high-resolution Mel-Frequency Cepstral Coefficients (MFCC) with online Cepstral Mean Normalization (CMN). The features are extended with online i-vectors [36, 37].

We used a vocabulary of 28.4k unique tokens. Out of this, 15.3k tokens are 5-letter way-points from [38], and 5.2k tokens are airline designators for call-signs<sup>6,7</sup>. Pronunciations were generated using Phonetisaurus tool [39] with a G2P model trained on Librispeech lexicon [40]. Standard 3-gram Language Model (LM) was obtained by interpolating LMs from individual training corpora. The ATC speech should follow the standard [41], but “free speech” can also appear in the data.

The English ASR achieved 8.4% Word Error Rate (WER) on our custom Airbus-dev dataset [13] and 8.9% WER on MALORCA project – Vienna airport test set [18]. The ASR achieved 20.4% WER on 5 minutes English subset of our ENCZ-Eval data<sup>8</sup>

To the best of our knowledge, there exists no Czech ATC speech corpus for training an in-domain ASR. Therefore, we relied on an off-the-shelf Czech ASR, which is 8kHz ASR system trained on a mixture of 2512 hours of telephone and distant microphone speech, including data augmentation. We used a hybrid speech recognizer trained with Kaldi. We used TDNN architecture [42], containing 7 layers each with 450 neurons. The NN was trained with Lattice-free MMI objective, sub-sampling by factor of 3 and bi-phone targets as suggested in [34]. The features were 40-dimensional Mel-filter bank outputs. We added only the local way-points, the airline designators for call-signs, and a hand-filtered set of alphabet and commands [41] to the vocabulary. This word list was mixed with the original LM with an ad-hoc 0.3 weight. The ASR achieved 19.1% WER on prompted Czech speech recorded on distant microphone in noisy environment and 30.7% WER on telephone speech conversations.

The ASR achieved 79.2% WER on 2 minutes of Czech subset from our ENCZ-Eval data. The high WER shows that the ASR is not adapted to the ATC domain. Since we are in the beginning of the transcription process, we have only 2 minutes of transcribed data so far. But, we can see that Czech ASR system has much higher WER (79.2%) compared to the English ASR system (20.4%) on our ATC data.

A more detailed description of the ASR systems is out of the scope of this paper. The reader is kindly asked to find the

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_airline\\_codes](https://en.wikipedia.org/wiki/List_of_airline_codes)

<sup>7</sup>[https://www.faa.gov/documentLibrary/media/Order/7340.2J\\_Chg\\_1\\_dtd\\_10\\_10\\_19.pdf](https://www.faa.gov/documentLibrary/media/Order/7340.2J_Chg_1_dtd_10_10_19.pdf)

<sup>8</sup>We are on the beginning of the transcription process so we have only 5 minutes of transcribed data so far. The provided WER cannot be taken as very reliable. On the other hand, it provides an idea about how well the ASR works.

Table 3: *Hyper-parameters of Bayesian SMM. so stands for semi-orthogonal constraint.*

Hyper-parameter	set / range
regularization type	$\{\ell_1, \ell_2, \text{so}\}$
regularization weight	$\{1e-05, 2e-05, \dots, 1e-02\}$
subspace dimension	$\{32, 64, \dots, 160\}$

details in [11].

## 5. ASR-based English Language Detection

As described in the Sections 1 and 2, the spoken utterances are code-mixed, i.e., each utterance may contain a greeting in one language (e.g. Czech) followed by a message in a different language (e.g. English). We hypothesize that the spoken words in the message context could be helpful in determining the *significant* language of the utterance. Hence we employed models that are prominent in text modelling / classification.

For each utterance (or recording), we extract bag-of-words statistics from the transcriptions automatically generated by the ASR. More specifically, we use both English and Czech ASR systems to generate confusion networks for every recording (confusion network is a sausage-like structure with bins containing soft counts or probabilities for a list of candidate words in each bin). The word statistics for each utterance are accumulated to form an utterance-by-word count matrix. The vocabulary sizes of EN and CZ are 27675 and 326509 respectively. We filtered out words whose total occurrence in training set is less than  $1e-03$ , which resulted in 149673 unique words. We can observe that the vocabulary size is much higher than the number of utterances (Table 1).

The filtered utterance-by-word counts are used by the Bayesian subspace multinomial model (BaySMM) [43] to learn utterance embeddings<sup>9</sup>. The model is trained in an unsupervised way, and can learn to represent utterance embeddings ( $\mathbf{u}$ ) in the form of Gaussian distributions, thereby encoding the uncertainty in the covariance matrix, i.e.,  $p(\mathbf{u} | \boldsymbol{\nu}, \text{diag}(\boldsymbol{\gamma})^{-1})$ , where  $\boldsymbol{\nu}$  represents the mean of the utterance embedding and  $\boldsymbol{\gamma}$  represents the diagonal precision (or the uncertainty). The estimated uncertainties are high for shorter utterances with more unique words as compared to longer utterances with many tokens. Modelling the uncertainties helps in training a robust classifier for downstream tasks.

The primary classifier for ELD is the generative Gaussian linear classifier with uncertainty (GLCU) [43, 44], which is trained on the utterance embeddings  $p(\mathbf{u})$  obtained from BaySMM. Additionally, we also train a binary logistic regression for comparison.

Our baseline ASR-based ELD system relies on *term frequency inverse document frequency* (TF-IDF) representation [45] of utterance statistics followed by a binary logistic regression for English / non-English classification.

## 6. Experiments and results

Since our aim is to filter-out non-English and retain as much English data as possible, we chose equal-error-rate (EER) as our evaluation metric.

We used ENCZ-Dev from Table 1 and ENIS-Dev from Table 2 as the training sets for unsupervised learning in BaySMM. Unless and otherwise mentioned, the utterance-by-word matrix

<sup>9</sup><https://github.com/BUTSpeechFIT/BaySMM>

Table 4: EER results on Eval sets using acoustic and ASR-based approaches. Training data are combined ENCZ-Dev and ENIS-Dev sets. dim refers to dimension of feature representation for the classifier (CLF).

Model	dim	CLF	ENCZ-Eval	ENIS-Eval
x-vector	256	GLC	0.0870	0.1450
BaySMM	64	GLCU	<b>0.0439</b>	<b>0.1003</b>
BaySMM	160	LR	0.0467	<b>0.1003</b>
TFIDF	149673	LR	0.0538	0.1084

Table 5: EER results on Eval sets using acoustic and ASR-based approaches. Training data is ENCZ-Dev only.

Model	dim	CLF	ENCZ-Eval	ENIS-Eval
x-vector	256	GLC	0.0890	0.2030
BaySMM	64	GLCU	0.0435	<b>0.1352</b>
BaySMM	128	LR	<b>0.0423</b>	0.1807
TFIDF	108000	LR	0.0554	0.1888

is obtained by pooling word statistics from the confusion networks of both EN and CZ ASR systems. No additional, external data was used in training the BaySMM. The embeddings are extracted for all utterances from all the sets. We performed 5-fold cross-validation classification experiments on the training set to choose the best model, i.e., hyper-parameter configuration (Table 3). The training details of x-vector based system is described in Section 3.

Table 4 presents the EER results for the baseline acoustic ELD based on x-vectors, and the proposed ASR-based ELD systems. The training data for classifier is the union of ENCZ-Dev and ENIS-Dev sets and the evaluation is done independently on ENCZ-Eval and ENIS-Eval sets. From Table 4, we see that all the ASR-based approaches outperform acoustic x-vector system. Further, the BaySMM has an EER of 0.0439, a 50% relative reduction as compared to the x-vector system. The major gains in ASR-based system are due to fact that the channel variability/noise in the data is addressed by the ASR and hence text embeddings and the classifier are not affected. On the other hand, the acoustic ELD has a more difficult task of dealing with the channel variability / noise, while aiming to classifying the acoustic features to the target language.

### 6.1. Effect of training data from target language / airport

Table 5 presents the results for an out-of-domain scenario, i.e., the classifier is trained only on ENCZ-Dev, and we evaluate on ENCZ-Eval and ENIS-Eval sets. This gives us an idea of how well systems perform on data from an unseen language / airport. We can see from the Table 5 that ASR-based approaches achieve lower EERs as compared to the acoustic system. The secondary language / airport from ENIS-Eval set is totally unseen in the training data, and we can see that ASR-based system has reduced the EER by 35% relative compared to the acoustic ELD. Having training data from target airport (ENIS-Dev) reduces the EER from 0.1352 to 0.1003, a 25% relative reduction.

### 6.2. English vs target-language ASR systems

We also analyse the effect of individual ASR systems for the ASR-based ELD. Table 6 shows the EER results for English and Czech ASR-based ELD systems. We can see (column ENCZ-Eval) that having only a matching secondary language (CZ) ASR helps to reduce the EER (0.0905  $\rightarrow$  0.0621) as com-

Table 6: EER results on Eval sets using individual ASR-based approaches. Training data are combined ENCZ-Dev and ENIS-Dev sets.

ASR	Model	dim	CLF	ENCZ-Eval	ENIS-Eval
EN	BaySMM	64	GLCU	0.0937	0.1352
	BaySMM	160	LR	<b>0.0905</b>	<b>0.1096</b>
	TFIDF	27675	LR	0.1008	0.1165
CZ	BaySMM	64	GLCU	<b>0.0621</b>	0.2331
	BaySMM	128	LR	0.0621	<b>0.1807</b>
	TFIDF	130262	LR	0.0850	0.2331

pared to having only an English ASR. However, having only a mismatched (CZ) ASR (column ENIS-Eval) increases the EER (0.1096  $\rightarrow$  0.1807), as compared to having only an English ASR. We plan to conduct a more detailed analysis as we collect more data from multiple airports / languages.

## 7. Conclusions

We presented our continuing efforts in collecting ATC speech data, which holds future for in-cockpit voice enabled applications. We described the need for filtering non-English speech, an important pre-processing step for downstream applications relying on the ATC speech data. We explored both acoustic and ASR-based approaches for detecting English utterances from the collected ATC data. Our results show that ASR-based English Language Detection (ELD) systems are robust and achieve low equal-error-rates as compared to acoustic ELD systems based on x-vectors. We have also shown that the ASR-based ELD can be applied to the new airport with a different secondary language, and achieves 0.1352 EER. This can be of great help as one can filter out non-English with minimal resources and an off-the-shelf secondary ASR. Moreover, the ASR-based ELD only requires to have a robust English ASR (secondary ASR gives additional improvements), whereas the acoustic ELD needs to be trained on multiple languages under consideration. Furthermore, ASR is a major component in downstream applications on ATC speech data, hence using the same ASR output for filtering out non-English is more suitable and efficient. We note that both acoustic ELD and ASR systems are data hungry on their own, whereas the embeddings from ASR and the classifier are not. Hence, retraining or adapting a text embedding model such as Bayesian subspace multinomial model will be relatively fast.

In future, we plan to collect more data from more languages and airports. More data will also enable experimenting with fusion of both approaches. As part of the ATCO<sup>2</sup> project, we plan to make the data public. We also plan to test our approach on the ATCSpeech Chinese / English speech data if available.

## 8. Acknowledgements

The work was supported by European Union’s Horizon 2020 projects No. 864702 - ATCO2 and No. 884287 HAAWAI. Part of high-performance computation run on IT4I supercomputer and was supported by the Ministry of Education, Youth and Sports of the Czech Republic through e-INFRA CZ (ID:90140).

## 9. References

- [1] Y. Lin, “Spoken instruction understanding in air traffic control: Challenge, technique, and application,” *Aerospace*, vol. 8, no. 3, 2021.
- [2] D. G. Martínez, O. Pichot, L. Burget, O. Glembek, and P. Matějka,

- “Language Recognition in iVectors Space,” in *Proc. of Interspeech*, vol. 2011, no. 8. ISCA, 2011, pp. 861–864.
- [3] A. Lozano-Diez, O. Plchot, P. Matejka, and J. Gonzalez-Rodriguez, “DNN Based Embeddings for Language Recognition,” in *IEEE ICASSP*, 2018, pp. 5184–5188.
  - [4] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, “Spoken Language Recognition using X-vectors,” in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 06 2018, pp. 105–111.
  - [5] C. Chandak *et al.*, “Streaming Language Identification using Combination of Acoustic Representations and ASR Hypotheses,” 2020.
  - [6] M. A. Zissman, “Language identification using phoneme recognition and phonotactic language modeling,” in *IEEE ICASSP*, vol. 5, 1995, pp. 3503–3506 vol.5.
  - [7] H. S. Lee, Y. Tsao, S. K. Jeng, and H. M. Wang, “Subspace-based representation and learning for phonotactic spoken language recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3065–3079, 2020.
  - [8] M. Soufifar, M. Kockmann, L. Burget *et al.*, “iVector Approach to Phonotactic Language Recognition,” in *Proc. of Interspeech*. Florence, Italy: ISCA, Aug. 2011, pp. 2913–2916.
  - [9] M. Soufifar, L. Burget, O. Plchot, S. Cumani, and J. Černocký, “Regularized subspace n-gram model for phonotactic ivector extraction,” in *Proc. of Interspeech*, Aug. 2013, pp. 74–78.
  - [10] S. Wang, L. Wan, Y. Yu, and I. L. Moreno, “Signal combination for language identification,” 2019.
  - [11] M. Kocour, K. Veselý, A. Blatt, J. Z. Gomez, J. H. Černocký, D. Klakow, and P. Motlíček, “Boosting of contextual information in ASR for air-traffic call-sign recognition,” in *submitted to Interspeech 2021*. ISCA, 2021.
  - [12] B. Yang, X. long Tan, Z. Chen, B. Wang, D. Li, Z. Yang, X. Wu, and Y. Lin, “ATCSpeech: a multilingual pilot-controller speech corpus from real Air Traffic Control environment,” in *Proc. of Interspeech*, 2020, pp. 399–403.
  - [13] E. Delpech, M. Laignelet, C. Pimm, C. Raynal, M. Trzos, A. Arnold, and D. Pronto, “A Real-life, French-accented Corpus of Air Traffic Control Communications,” in *Proc. of LREC*, 2018.
  - [14] J. Godfrey, “The Air Traffic Control Corpus (ATC0) - LDC94S14A,” <https://catalog.ldc.upenn.edu/LDC94S14A>, 1994.
  - [15] K. Hofbauer, S. Petrik, and H. Hering, “The atosim corpus of non-prompted clean air traffic control speech,” in *LREC*, 2008.
  - [16] L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, “Design and development of speech corpora for air traffic control training,” in *Proc. of LREC*. European Language Resources Association (ELRA), May 2018.
  - [17] L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, “Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development,” *LREC*, vol. 53, no. 3, pp. 449–464, 2019.
  - [18] 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 Interspeech*. ISCA, Aug. 2017.
  - [19] C. Swail, L. Benarousse, E. Geoffrois, J. Grieco, R. Series, H. Steeneken, H. Stumpf, and D. Thiel, “The NATO Native and Non-Native (N4) Speech Corpus,” 01 2003.
  - [20] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Proc. of Interspeech*. ISCA, Aug. 2017.
  - [21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” *IEEE ICASSP*, pp. 5329–5333, 2018.
  - [22] O. Novotný, O. Plchot, P. Matějka, L. Mošner, and O. Glembek, “On the use of x-vectors for robust speaker recognition,” 06 2018, pp. 168–175.
  - [23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE CVPR*, 2016, pp. 770–778.
  - [24] “The 2009 NIST Language Recognition Evaluation (LRE09),” <https://catalog.ldc.upenn.edu/LDC2014S06>.
  - [25] “The 2011 NIST Language Recognition Evaluation (LRE11),” <https://www.nist.gov/itl/iad/mig/2011-language-recognition-evaluation>.
  - [26] “The 2015 NIST Language Recognition Evaluation (LRE15),” <https://www.nist.gov/itl/iad/mig/2015-language-recognition-evaluation>.
  - [27] “The 2017 NIST Language Recognition Evaluation (LRE17),” <https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation>.
  - [28] P. Roach *et al.*, “Babel: An eastern european multi-language database,” in *Proc. of ICSLP*, vol. 3, 01 1996.
  - [29] “The 2016 NIST Speaker Recognition Evaluation Plan (SRE16),” <https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016>.
  - [30] C. Cieri, D. Miller, and K. Walker, “The Fisher corpus: A resource for the next generations of speech-to-text,” Jan 2004.
  - [31] D. Povey *et al.*, “The Kaldi Speech Recognition Toolkit,” in *IEEE Workshop on ASRU*. IEEE Signal Processing Society, Dec. 2011.
  - [32] O. Novotný, O. Plchot, O. Glembek, J. Černocký, and L. Burget, “Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition,” *Computer Speech and Language*, vol. 2019, no. 58, pp. 403–421, 2019.
  - [33] J. Segura *et al.*, “The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication,” 2007.
  - [34] D. Povey *et al.*, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in *Proc. of Interspeech*. ISCA, Sep 2016, pp. 2751–2755.
  - [35] —, “Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks,” in *Proc. of Interspeech*, 09 2018, pp. 3743–3747.
  - [36] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, “JHU aspire system: Robust LVCSR with tdnns, ivector adaptation and RNN-LMS,” in *Workshop on ASRU*. IEEE, Dec 2015, pp. 539–546.
  - [37] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in *Workshop on ASRU*. IEEE, Dec 2013, pp. 55–59.
  - [38] X. Olive, “Traffic, a toolbox for processing and analysing air traffic data,” *Journal of Open Source Software*, vol. 4, p. 1518, 2019.
  - [39] J. R. Novak, N. Minematsu, and K. Hirose, “Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework,” *Nat. Lang. Eng.*, vol. 22, no. 6, pp. 907–938, 2016.
  - [40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *ICASSP*. IEEE, 2015, pp. 5206–5210.
  - [41] “Aeronautical Telecommunications, Annex 10, Volume II,” ser. Edition 6. International Civil Aviation Organization, Oct 2001.
  - [42] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. of Interspeech*, ISCA. ISCA, Sep 2015, pp. 3214–3218.
  - [43] S. Kesiraju, O. Plchot, L. Burget, and V. S. Gangashetty, “Learning document embeddings along with their uncertainties,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 2020, no. 28, pp. 2319–2332, 2020.
  - [44] S. Cumani, O. Plchot, and R. Fér, “Exploiting i-vector posterior covariances for short-duration language recognition,” in *Proc. of INTERSPEECH*, no. 09, Dresden, Germany, 2015, pp. 1002–1006.
  - [45] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 2001.