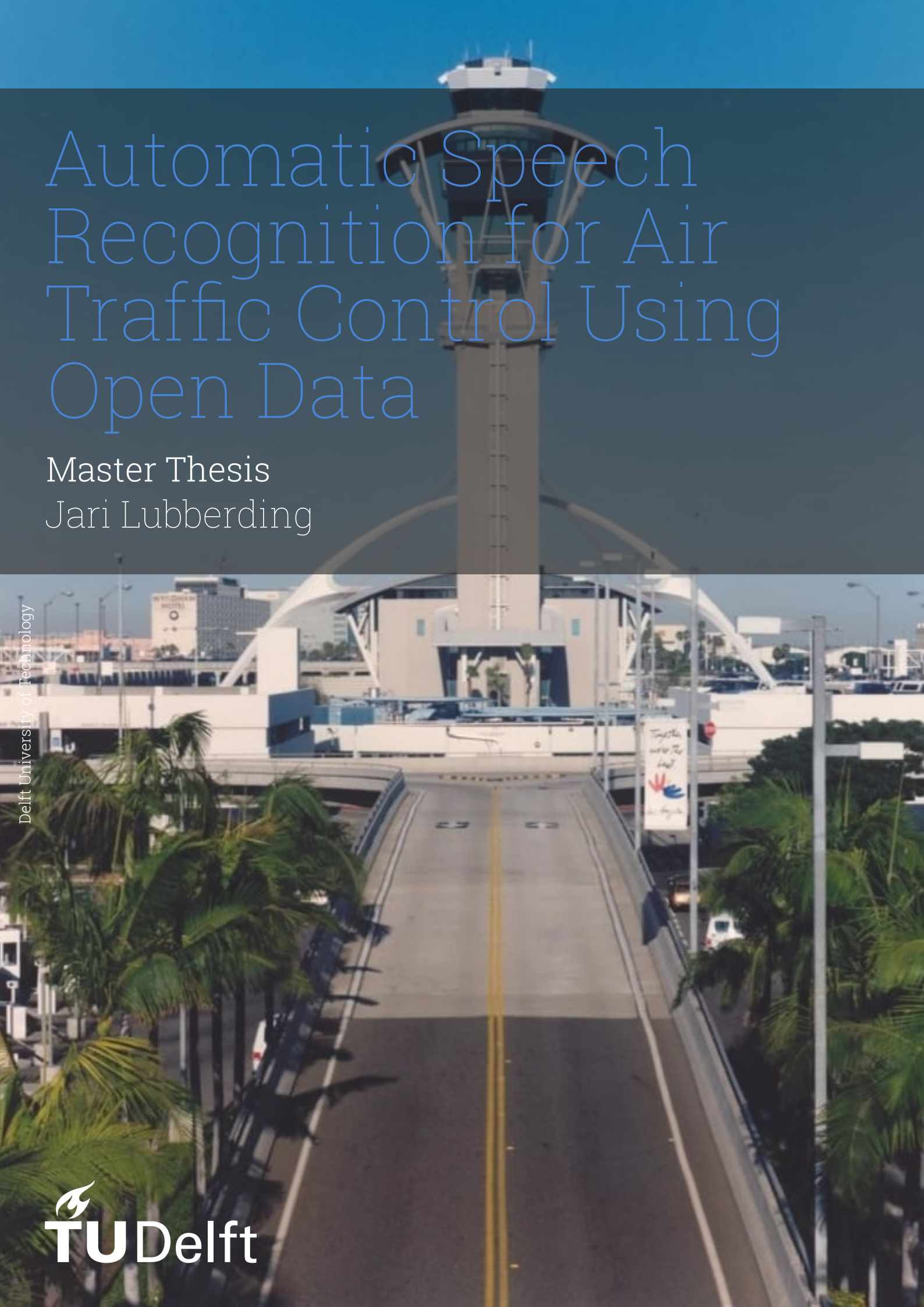
Automatic Speech Recognition for Air Traffic Control Using Open Data

Master Thesis

Jari Lubberding



Delft University of Technology

TU Delft

Automatic Speech Recognition for Air Traffic Control Using Open Data

by

Jari Lubberding

Student Number

4291867

Thesis Committee:	Dr. J. Sun	Responsible Thesis Supervisor
	Prof. Dr Ir. J.M. Hoekstra	Committee Chair
	Dr. X. Wang	Examiner
Graduation Date:	28-08-2023	
Faculty:	Faculty of Aerospace Engineering, Delft	

Cover: LAX Air Traffic Control Tower by Swinerton

TU Delft

Automatic Speech Recognition for Air Traffic Control Using Open Data

Jari Lubberding
Supervised by: Junzi Sun, Jacco Hoekstra
Control and Simulation, Faculty of Aerospace Engineering, TU Delft

Air Traffic Control (ATC) is tasked with ensuring safe separation between aircraft in a given Controlled Traffic Region (CTR). To achieve this an Air Traffic Controller (ATCo) verbally gives clearances using over the air communication. These clearances are kept track of by the ATCo using so-called 'flight-strips', which in modern systems are often digital. The allocation of an ATCo's time is an important factor in the achievable traffic density within a CTA, which makes ATC an interesting domain to use Automatic Speech Recognition (ASR) models to allow a computer system to 'listen in' to the conversation of the ATCo. Although previous research has been done to create such models, few of these result in open available models or domain specific corpora for the creation of such a model. This study will therefore use two open in-domain and one out-of-domain corpora to create such a model and in this process identify domain specific challenges and how these challenges can, in certain cases, be mitigated.

I. Introduction

The main role of Air Traffic Control (ATC) is to ensure safe separation between aircraft in a given Controlled Traffic Region (CTR). This is achieved by giving clearances to each aircraft over radio-communications which coordinates the traffic to ensure safe separation. These clearances are traditionally kept track of by the Air Traffic Controller (ATCo) on so-called 'flight-strips' and denote the clearances given to a certain aircraft, these can then easily be passed on to other ATCos in charge of different Control Areas (CTAs) which are managed within the same physical room. Modern controller working positions often implement digital flight-strips, which have the advantage that they can be passed on to different controllers which are not located in the same location. Additionally they allow the clearances to be integrated within the radar screen, allowing for a more information rich Human Machine Interface (HMI). However, this still requires the ATCo to manually enter the clearances given to a certain aircraft, even-though the information is simply a repetition of the what the controller has told the pilot. As the allocation of the ATCo's time is an important factor in the achievable traffic density within a CTA, it is important to use an ATCo's time as efficiently as possible. This has lead to several applications of Automatic Speech Recognition (ASR) within ATC such as:

- The AcListant® project (developed by Saarland Universit and DLR) which has shown great results with respect to ATCo workload reduction, which uses ASR to automatically maintain the clearances given to aircraft on the digital flight-strips[1].
- The usage of ASR in ATCo training simulators due to the large cost associated with the training an ATCo. These costs are largely incurred due to the need of so called *pseudo-pilots*. These pseudo-pilots are often retired pilots who will act as virtual planes in training scenarios. Šmídl et al. present an approach to use both ASR and Text To Speech (TTS) models to replace human pseudo-pilots with a computer in an effort to reduce to cost of training an ATCo[2].
- Cordero et al. present a way to use ASR to automatically and objectively estimate the workload of an ATCo. Where traditional approaches heavily relied on inferring workload based off ATCo interactions with the user interface, this new approach allows the content of the communications to be used to more accurately infer ATCo workload. This enables dynamical load-balancing of ATCo during operation and can give a more objective metric of ATCo workload

during experiments[3].

- Chen et al. present a system that uses ASR to detect if an ATCo makes use of a closed runway and thereby increasing the surface safety of the airport. More recently [4] present a system to automatically detect pilot read back errors that are not corrected by the ATCo, greatly reducing the risk of miscommunications between the ATCo and pilot[5].

These applications make research into ASR models capable of extracting command information in the domain of ATC an interesting topic. Access to such a model could open up a wide variety of new research opportunities such as the in-depth characterization of an airspace without direct access to proprietary data. However unfortunately few of the research conducted leads to openly accessible models or domain specific data-sets required for the creation of such a model.

Having access to such a model would be beneficial for research associated with the Air Traffic Management (ATM) department within the faculty of Aerospace Engineering (TU Delft). The goal of this thesis is, therefore, to create an ASR model using openly available corpora and tool-kits. This process will be useful for identifying the domain specific challenges, whether these challenges can be mitigated, and ultimately to make recommendations what areas are worthwhile for future research.

II. Related Works

This section will discuss related work with the primary goal of identifying the necessary elements of an ASR model capable of the extraction of callsign and command data. The first section will explain the basic working principles of an ASR model and will discuss different state-of-the-art modeling approaches. The second section will discuss the different corpora, or ASR specific datasets, available. Finally the methods employed by other researchers to incorporate context information and extract callsigns and commands will be presented.

A. Automatic Speech Recognition

The assumption made by ASR models is that the acoustic signal is a realisation of a sequence of symbols. To decode this sequence of acoustic realisations of symbols these models first convert the audio signal into a sequence of equally spaced acoustic parameter vectors in a process referred to as feature extraction. The recognizer is then tasked with the extraction of the original sequence of symbols from the extracted features. For the recognition of an isolated word the problem can be written as the maximum probability of a word w_i, within the total vocabulary, given the feature sequence $\mathbf{Y}$ as seen in Equation 1.

$$w = \arg\max_i \left[P(w_i|\mathbf{Y}) \right] \tag{1}$$

However this probability is hard to determine directly and therefore it is rewritten using Bayes' Rule into Equation 2 and consists of three likelihoods. The first $P(\mathbf{Y}|w_i)$ is called the Acoustic Model (AM) as it models the likelihood of a certain word given the acoustic input signal. The second $P(w_i)$ is referred to as the Language Model (LM) and gives the probability for the occurrence the word itself, which can for example include the knowledge of the previous word to improve recognition in so called n-gram models. The final one $P(\mathbf{Y})$ is irrelevant as the sequence $\mathbf{Y}$ is a given and is therefore usually omitted.

$$P(w_i|\mathbf{Y}) = \frac{P(\mathbf{Y}|w_i)P(w_i)}{P(\mathbf{Y})} \tag{2}$$

The previous example assumes single word recognition where there exists a model for each word in the entire vocabulary that tries to estimate the likelihood of the feature sequence to be the result of that specific word. However, many words consist of similar sounds and hence smaller speech units, such as phones, are generally used to model speech. Practically this means that the acoustic model tries to estimate the probability of a single phone, or tri-phones when the phone is assumed to sound different depending on neighbouring phones. These phones are then *'strung'* together to achieve word and sentence recognition, often referred to as *'beads-on-a-string model'*. The usage of phones as a smaller speech unit requires the use of a 'lexicon' which maps words to their respective sequence of phones. The more generic version of Equation 2 excluding $P(\mathbf{Y})$, where $\mathbf{W}$ is the word sequence, is given as:

$$P(\mathbf{W}|\mathbf{Y}) = P(\mathbf{Y}|\mathbf{W})P(\mathbf{W}) \tag{3}$$

The ASR research made major steps towards usable systems throughout the 1980s to 2000s with the

use of so called Hidden Markov Models (HMMs) statistical methods of modeling speech dynamics in continuous speech recognition models[6]. These often used Guassian Mixture Models(GMM) to describe the observation distribution of the HMMs, these were then referred to as GMM-HMM. However now Deep Neural Networks (DNNs) have gotten a dominant role in ASR. The DNN based ASR models can be divided into two main categories, the first are the DNN-HMM hybrids and the second are the End to End (E2E) models. DNNs are very capable when used as static classifiers where the inputs have a fixed dimension. However, the application ASR is a sequential recognition problem. This therefore led to the development of hybrid models that combined the static classification strength of a DNN with the HMM strength of handling sequential patterns[7]. These models, however, are very cumbersome to train because a bootstrap model is required. This means that first a GMM-HMM has to be trained to obtain frame-level alignments which is then used to train the DNN-HMM. These systems are very complex and has given rise to recent development in E2E models. These models do not need a bootstrap GMM-HMM or a phonetic decision tree, can be trained in a single stage, can remove the need of a pronunciation lexicon and finally the AM and LM can be trained jointly. Although training the AM and LM jointly can seem appealing, these methods are very data hungry needing thousands of hours of training data to be competitive with the hybrid models[8]. CAT(CTC-CRF ASR Toolkit) is an interesting new toolkit that has been developed which allows for E2E training of an AM while still being able to use a separate LM, therefore still being very data efficient but eliminating the cumbersome training process of a DNN-HMM[8].

B. Corpora

ATC has a clear advantage over regular speech as the ICAO standard phraseology has a very limited vocabulary [9], this means that a ASR model has to recognise fewer words. Additionally, it also has a different way of pronunciation certain words. This is done to make the communications such that they can be interpreted over the noisy Very High Frequency (VHF) transmission medium. The consequence is that conventional corpora are not sufficient to train an ASR model for the ATC domain because the ATC domain contains pronunciations unknown to these conventional corpora.

The noise due to the transmission medium is an additional challenge, although not necessarily when the ASR model is only tasked with listening to the ATCo as it can be recorded directly at the source. However, such data is often hard to obtain. As mentioned by Hofbauer et al. an Air Traffic Control Centre (ATCC) often records all radio communications between ATCo and pilots for legal reasons, however publicly sharing this data would be legally problematic in most European countries[10]. Most corpora relied on the access to such data, where Šmídl et al. [2] used industrial partners to gain access to such recordings, whereas Hofbauer et al. relied on resources to simulate a realistic scenario for their recordings[10]. However, more recent research by [11] showed the use of *liveatc.net* as a source of such recordings. *liveatc.net* is a site which provides access to live recordings of many different airspace regions which are recorded by hobbyists and as described by Zuluaga-gomez et al. "can be considered as an 'unlimited-source' of low-quality data."[11]

As previously mentioned the use of domain specific corpora is important for the performance of an ASR model. Currently there are five domain specific corpora publicly available, these are:

- The NIST - Air Traffic Control Complete. Available through the Linguistic Data Consortium (LDC) for a fee depending on membership status[12].
- The non-native Military Air Traffic Communications (nnMATC) database. The dataset is NATO unclassified and its use is restricted to language and speech research. To gain access to it, researchers must contact the first author of the paper presenting the dataset[13].
- The HIWIRE corpus was created by to enable the improvement of ATC ASR models with respect to noisy and non-native speech. Although the paper mentions the database is freely available, it can only be obtained through a paid licence from elra[14, 15].
- The ATCOSIM dataset "provides 50 hours of publicly available direct-microphone recordings of operational air traffic controller speech in a realistic civil en-route control situation." This

corpus is publicly available and free of charge[10, 16].

- The Air Traffic Control Communication (ATCC) created by Šmídl et al. uses actual ATC communication recordings, which has significant level of noise in the signal, non-native English accents of the speakers, non-standard pronunciation of some frequent words and a rather limited vocabulary. The data set is publicly available and free of charge[17, 18].

The lack of open corpora available specific for this domain can be seen as only two of the previously five mentioned corpora are freely available.

C. Context Information and Command Extraction

The specific research of ASR within ATC primarily focuses on the inclusion of context information to increase recognition rates and methods of extracting callsign and command data. This context can be for example all the current aircraft within a certain airspace region. One of the earlier research projects is the AcListant® led by a collaboration between the University of Saarland and the German Aerospace Center (DLR) did initial research on lattice re-scoring using context information[19]. The paper uses a grammar-based LM, based on the ICAO phraseology[9], to penalize certain hypothesis generated by the model. Although the previous method worked, it was very computationally expensive. In addition to that, creating the rules to describe the grammar which accurately represented the utterances was difficult due to the tendency of ATCos to deviate from standard phraseology[20].

This led Oualil et al. away from the grammar-based LM and instead chose to adopt an statistical LM, more specifically a trigram LM, which outperformed the rule-based grammar.[20] Instead of incorporating the contextual information into the LM, it was now chosen to implement it in a post-processing step. The trigram LM was trained using both labeled data and generated data from the grammar model. A big disadvantage from moving away from a predefined grammar, is that the transcriptions are no longer in a predefined format and therefore has become harder to directly extract the semantic information. To now extract the semantic information the researchers used a sequence labeller. Both tested sequence labellers, being Conditional Random Field (CRF)-based tagger and Context Free Grammar (CFG)-based token tagger, performed equally well.

This briefly shows that the field of ATC moved away from trying to incorporate context information directly into the ASR model. It instead incorporates this information as a post processing step using the multiple hypothesis generated by an ASR model.

III. Methodology

This section will first give a description of available data and each of these corpora will be be discussed, giving a general overview of their merits. The second section will describe the data processing steps that were taken to prepare the data for training. This will reveal how design decisions of these corpora influence their usefulness.

A. Data Description

In Section II it was shown that openly available in domain corpora are scarce. This section will describe these various corpora.

1. ATCOSIM

Although the previously mentioned "50 hours of publicly available" data is only around 10.7 hours of actual speech. It includes a dictionary of all occurring words, however it does not include a phonetic dictionary mapping words to their respective phones. The recordings are created from a simulated scenario and only contain the recordings of the Air Traffic Controller (ATCo). The sound is therefore high quality and can be used to evaluate the potential performance of an ASR model which operates directly at the location of recording. It has a relatively limited set of words of around 800 and little to no deviation from standard phraseology. This makes this corpus not realistic on the basis of language complexity. The recordings consist of 10 different individuals of both sexes with several different accents. This makes this corpus useful for evaluating the influence of accents on performance. Each utterance is its own recording and all of them are transcribed word-for-word in standard British English. To give an illustration of a transcription:

topswiss four five seven eight climb to flight level three one zero

2. Air Traffic Control Communication (ATCC)
The corpus is created by Šmídl et al. and in the paper presenting the corpus a model is trained on 140 hours of data[17, 18]. However the corpus that is publicly available consists of 20.6 hours of raw recordings of which 10.6 hours is actual speech. The corpus consists of recordings taken from airspace in the Czech Republic. This corpus is therefore a more realistic application of such an ASR model. However due to the nature of the recording there is no information on the speakers. This means that during training it cannot be guaranteed that a speaker is not both in the training as in the evaluation set, potentially causing what is called cross contamination. The data structure is different to ATCOSIM. Instead of each utterance having its own sound file, this corpus has an xml file associated with each sound file in which multiple subsections are marked and transcribed as shown here:

```
<Sync time="13.460"/>
   [air]head 0 8 0 Easy (9(najn)) C L
<Sync time="15.690"/>
```

This also shows other characteristics of this corpus. First is that the utterance is labeled with [air], a similar tag exists for ground. This makes it possible to evaluate the performance of a model with respect to air and ground separately. Additionally it can be seen that the transcription method employed differs from that of ATCOSIM. Instead of using word-for-word the designers of this corpus opted to instead transcribe it more from the perspective of an ATCo. The C and L for example are pronounced as Charlie and Lima respectively. In contrast to ATCOSIM this corpus does include a phonetic dictionary, which according to the paper uses the Arpabet transcription code, consistent with CMU Dict. Finally, the authors have also trained their own model which allows for a comparison to the results presented in this paper. The training method employed by Šmídl et al. is to first train a model on both ground and air, and to then retrain the base model with only data from each respective entity.

3. LibriSpeech
LibirSpeech is selected as an out of domain corpus to evaluate whether increasing the amount of data from other sources can improve recognition rates. LibriSpeech consist of around of 1000 hours of read speech from books. The 1000 hours are separated into two 'clean' train sets of each 100 and 360 respectively and a 'other' train set of 500 hours and comes with a validation and test set for both 'clean' and 'other'. It includes a lexicon with phonetic definitions of the words based on CMUDict. The set is chosen because it is freely available and often used as a standard benchmark test for ASR models.

B. Data Preparation
The data preparation for all corpora will be described in this section, this will include both the choices made with respect to certain tags within the transcriptions and the creation of the phonetic lexicon. The model training will be done in the ASR toolkit Kaldi. Kaldi requires the creation of a certain set of files with a particular structure such that it can be used its tools. These files are:

- *spk2gender*: Maps a gender to a specific speaker
- *wav.scp*: Connects every utterance with an audio file related to this utterance
- *text*: Contains every utterance matched with its text transcription
- *utt2spk*: Tells the ASR system which utterance belongs to particular speaker
- *corpus.txt*: Contains transcriptions of all utterances
- *lexicon.txt*: Contains every word mapped to its individual phones
- *nonsilence_phones.txt*: All non-silent phones in the lexicon
- *silence_phones.txt*: All silent phones in the lexicon

To generate these files some decisions will be made for each corpus and will be discussed in their respective sections. Also, as mentioned before, words consist of similar sounds and hence these models are trained on the individual phones of a word. To transform the sequence of words into a sequence of phones a lexicon

is required which maps each word which occurs in the corpus to its corresponding phones. The word *Air* consists, according to CMUdict, of the following phones *EH1 R* where the 1 indicates lexical stress. The creation of this dictionary will also be discussed in the following sections.

1. ATCOSIM

Since the ATCOSIM corpus is created in a well controlled environment only minor changes have to be made. The corpus include tags for off-talk(<OT>), language that is not related to ATC, these tags encapsulate transcribed sections. Only the tags are removed and the utterances are kept. Several tags such as [UNKNOWN] and [HNOISE] are replaced with a single word <SPN> which will be mapped to what is called a garbage phone, a phone that is matched on all words that do not exist in the vocabulary. Utterances that contain the foreign language tag (<FL>) are removed since these are not transcribed, together with the [EMPTY] utterances.

Since ATCOSIM does not come with a phonetic lexicon, one has to be created. To achieve this first a list of phraseology specific words was created. These include words such as three and nine, which under phraseology are pronounced tree and niner respectively. However since the corpus is relatively small, the actual pronunciations were investigated and it turned out that nine was just pronounced as nine. So the regular phonetic, as per CMU Dict, were kept instead of the phraseology ones. Additionally other specially denoted 'words' such as l were manually added. The remainder of the words were then taken from CMU Dict, and the words that did not occur in the CMU Dict were generated using LOGIOS which is a rule based lexicon generation tool.

2. Air Traffic Control Communication (ATCC)

In the ATCC corpus the multiple utterances can exist in the same file and hence these need to be extracted. Additionally each of these utterances are labeled [ground], [air] or sometimes a combination of both in which a portion is encapsulated by ground and the other by air. Therefore the corpus was split up into three different groups *Air*, *Ground* and *Combined* wherein the latter has both the air and ground label in the utterance. Each of the utterances is typically associated with a specific speaker, however due to the nature of the data this information is not available. Therefore the assumption is made that within a single xml file associated with a specific sound file (average length of 30 seconds) that the ground speaker remains the same person, whereas each air or combined utterance is considered its own speaker. The tags [noise] and [speaker] were replaced with <NSN> which is a separate garbage phone for noise. Several tags such as [ehm ??] were replaced with <SPN> and many other tags were removed.

Although the corpus has a phonetic lexicon, it was not complete. In the first place not all words were actually present in the lexicon. This had to be expanded by using the CMU Dict. The words in the lexicon did not match the transcription, such as words enclosed in brackets having spaces between words in the transcription but not in the lexicon. Additionally the lexicon does not conform to the CMU Dict phoneset of 39 phones, instead this lexicon has 43 phones. This unfortunately makes it impracticable to augment this corpus with LibriSpeech, and also makes the additions which are made using CMU Dict not entirely consistent. There is repetition of the same word one with a capitalized letter and others without:

```
arrivals ah r ay v ah l z
Arrivals ah r ay v ah l z
arrivals ah r ay v l z
Arrivals ah r ay v l z
arrivals er ay v ah l z
Arrivals er ay v ah l z
```

It creates the impression that much of the lexicon is automatically generated and no proper data cleaning was done. Finally the lexicon also reveals a large disadvantage to transcription method used. Namely that it creates more ambiguity than necessary. For example *A* can now be the letter *A* or *Alpha*, two wildly different pronunciations which have to be implicitly learned by a model. Similarly the transcription of *1500* can be pronounced as *one thousand five hundred*, or as *fifteen hundred*. Which both occur in the lexicon, yet the entire point of ATC phraseology is to reduce ambiguity to a minimum and hence a word-to-word transcription seems much more appropriate.

To make the lexicon usable, a similar process as

for ATCOSIM was used. First all important specific phraseology and special characters were manually created, such as A B etc. (alpha bravo). All words were lowercased and merged and compared to their CMU Dict respective transcription and finally all missing words from the lexicon were then added from CMU Dict.

IV. Experimental Setup

The experimental setup will first describe the Automatic Speech Recognition (ASR) model structure. In the second section the evaluation metrics and the training procedure will be described. Finally the specific experiments will be described including a description on the specific question it attempts to answer.

A. ASR Model Description

As previously described the field of ASR research very active which has resulted in many different types of models. These models can be subdivided into two main categories, namely the End to End(E2E) and traditional models. The E2E models are trained directly from sound to transcribed text. These models therefore do not require an lexicon to map word to their respective sequence of phones. This greatly eases the process of training since there is no need for an AM and a separate LM. The disadvantage however is that these models are much more data hungry to train, as there is no way to inform the model of underlying relations between similar sounds. Due to their hungry nature this type of model is therefore inappropriate for this application as the amount of data is very limited. This therefore limits the selection to the traditional models which consist of an AM and a LM. However more recent hybrid approach try to combine newer E2E methods while still allowing the use of a LM to achieve state-of-the-art performance with a limited amount of training data while greatly decreasing the training complexity[8]. This study will use such a hybrid model as it greatly eases the process of training and reduces training time as it does not require an initial model to generate the alignments for training. This allows the training of such a model from around 6 hours (for a not augmented and no speed perturbed corpus, with around 10 hours of training data) to around 60 hours (for an augmented corpus, of around 110 hours of training data) on relatively modest hardware, being an Intel i5 2500k CPU and a Geforce GTX980Ti GPU.

The model is trained using conditional random field (CRF), specifically with connectionist temporal classification (CTC), or short CTC-CRF. The toolkit used is CAT (version 2)[8] which is an extension to the ASR toolkit Kaldi. The input feature to the model is an is a 80-dimensional fbank. Each utterance is normalized using cepstral mean and variance normalization (CMVN). The acoustic model is a 6 layer bidirectional LSTM with 320 hidden units as used by Xiang et al. to achieve state of the art performance with hybrid training[8].

The LM that will be used will be a regular n-gram, more specifically a 3-gram. An n-gram language model takes $n-1$ previous words into account in estimating the likelihood of a word. This means that in this particular case the language model will use up to two previous words to estimate the likelihood of the following word.

B. Evaluation Metric and Training Procedure

The *Word Error Rate (WER)* is a typical measure of performance for an ASR model. It calculates the percentage of incorrectly identified words. This is done using the following formula in which D is the amount of deletions, I is the amount of insertions and S is the amount substitutions which is divided by the total amount of words in the reference, meaning the summation of insertions (I), deletions (D) and correct words given by C:

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

The training procedure requires the corpus to be split up into three sets, being a train set (90%), a validation set (5%) and a test set (5%). The model will train using the train set, after every epoch (all data in train set has been used) the model will be evaluated using the validation set, once the loss of the validation set does not meaningful diminish the training is stopped. Finally the performance will be evaluated using the test set which has not been shown during training producing the WER presented in Section V.

C. Experiment Description

The following experiments will be explicitly demarcated in the results, however combined they intend to also answer other questions such as the benefits of increasing training data using speed perturbations and the performance of clean simulated data.

- **Influence of Accent:** This experiment will try to answer how important it is to have a model trained on a location specific corpus. To be more specific it will investigate the influence of an accent on the variance within the corpus itself. This will be tested by using the ATCOSIM corpus, which will be trained three times each using a different split of speakers. The first split, referred to as German Split 1, will use the only two German males as the validation (referred to in the corpus as gm2) and test set (referred to in the corpus as gm1), leaving no German males in the train set. The second split, referred to as German Split 2, will again leave no German males in the train set but it will reverse the males in the validation set (gm1) and test set (gm2), this is done to ensure that potential poor performance is not due to one particular speaker having a high variance with respect to the other speakers but that it is actually caused by its accent. The third split, referred to as Swiss Split, will use one Swiss male as the test set, leaving three Swiss males in the train set, hopefully creating an experiment in which there is a low variance between train the train and test set. In addition each experiment will be ran once with 3 fold speed perturbations, meaning the data gets sped up and slowed down with 10 percent to increase training data. This brings the total amount of models to be evaluated to six.
- **Out of Domain Data Augmentation:** To evaluate whether the inclusion of an out of domain corpus will improve recognition rates will be answered by comparing the performance of two of the previously three mentioned splits trained using three methods. The two splits are German Split 1 and Swiss Split, this can answer if a higher variance set benefits differently from that of a low variance set. The first of the three training methods is to only use the original split (same as the no speed perturbation from the Influence of Accent experiment). The second is run is the original split combined with the LibriSpeech 100 hours of clean data, to evaluate whether including out of domain data improves recognition rates. The third run uses the model weights of the second run to initialize the model and then to retrain the model only on the original split, to evaluate if bootstrapping a model (to use a model with initialized weights using additional data) produces any benefit.
- **Performance of Realistic Data:** This experiment will use the ATCC corpus to do four runs to investigate the performance on realistic data separated in air, ground and combined. The run trains the model using air, ground and combined together and evaluates them separately. The second does the same, but with speed perturbations. The third and forth are similar in that they retrain the second model each using only air and ground data respectively.

V. Results

The following section will show the results of the respective experiments. First however it is worth mentioning that the convergence graphs such as in Figure 1 are generated for all of the runs, in which the left figure shows the training loss and the right figure shows the loss on the validation set after each epoch. Most of these figures are not particularly interesting to show since figures between different experiments can only be directly compared when the same training set and validation set are used, which for the majority of the runs is not the case.

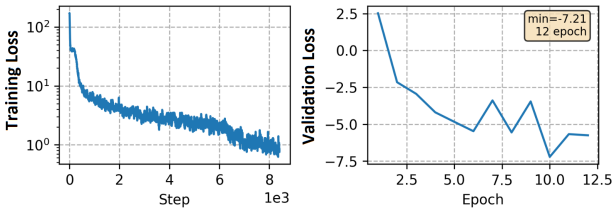


Figure 1. Model convergence of Swiss Split without speed perturbations. Left figure showing training loss and the right figure the loss on the validation set after each epoch.

A. Influence of Accent

Table 1 shows the Word Error Rates(WERs) of each of the different splits trained exclusively on the ATCOSIM corpus. The rows represent the different splits, where the both German Splits do not include the German speakers into the training data, only in the validation and test set. The Swiss Split uses one out of four Swiss speakers as test set and the remainder remain in the train set. The first column shows the result for training without speed perturbation, and the second shows the result for training with speed perturbation to increase training data.

ATCOSIM Corpus	No Speed Perturbation	Speed Perturbation
German Split 1	16.24	14.28
German Split 2	9.97	9.86
Swiss Split	4.28	4.38

Table 1. WERs in percentage for the three different splits of the ATCOSIM corpus.

B. Out of Domain Data Augmentation

Table 2 shows the WERs for the two splits. The first split is the German Split 1 where no German speakers are included in the training set and should therefore have a high variance with respect to the training set. The second split is the Swiss set which has one Swiss speaker as test set and three Swiss speakers in the train set and hence should have low variance with respect to the train set. The columns in table show the different augmentation methods, the first without augmentation, then with augmentation using LibriSpeech and finally the augmented model retrained using only the ATCOSIM split. Additionally Figure 2 shows the different convergence behaviour of Swiss Split when retraining when compared to the convergence behaviour of the original Swiss Split seen in Figure 1.

C. Performance of Realistic Data

Table 3 shows the WERs for the realistic data. The rows designate that the utterances are separated in Air, Ground and Combined fragments. The first column is the reference performance from the paper

ATCOSIM Corpus	No Speed Perturbation	LibriSpeech Augmented	LibriSpeech Augmented Retrained
German Split 1	16.24	11.21	14.06
Swiss Split	4.28	4.19	5.00

Table 2. WERs in percentage of the German Split 1 and Swiss Split from the ATCOSIM corpus. Showing in the columns without and with augmentation using LibriSpeech and retrained.

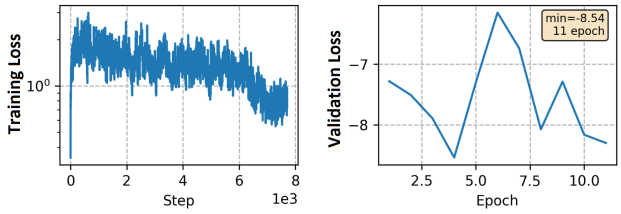


Figure 2. Model convergence of retraining of augmented model using Swiss split. Left figure showing training loss and the right figure the loss on the validation set after each epoch.

published by Smidl et al.. The following columns show the performance of the CTC-CRF model trained in this study, where the second column has no speed perturbation, the third column uses speed perturbation and the fourth column uses the model from the second column to initialize its weights and retrains on only Air and Ground respectively.

VI. Discussion

This section will discuss the results presented in Section V and put them into context of the intended goal of the experiments.

A. Influence of Accent

The goal of this experiment was to identify how important it is to have a model trained on a location specific corpus where it will be applied. To achieve this the German Splits completely removed the German accented speakers from the training set, whereas

ATCC Corpus	ATCC Paper	No Speed-Perturbation	Speed-Perturbation	Re-trained
Air	25.27	33.68	33.28	33.88
Ground	7.59	13.69	14.57	14.08
Combined	-	44.92	46.14	-
Average	-	27.02	27.60	-

Table 3. WERs in percentage of realistic data from the ATCC corpus. With in each respective column the ATCC paper results, no speed perturbation, speed perturbation and retrained with respect to Air and Ground

the Swiss Split left the majority of Swiss speakers in the training set. Table 1 shows that the performance of both the German Splits are considerably worse than that of the Swiss Split. Since both German Splits perform considerably worse it can be reasonably be attributed to their accent not being well represented in the training set, instead of just one particular speaker performing poorly. This shows that it is important to have an accurate representation of the domain in the corpus used when the corpus is particularly small(smaller than 10 hours). Table 1 also shows that it both the German Splits improve recognition rates when using speed perturbations, although German Split 2 only marginally. The well represented Swiss set, however, decreases slightly in performance. This indicates that speed perturbations might be useful in cases where test set is not well represented in the training set, whereas when the test set is well represented it does not meaningfully improve recognition rates.

B. Out of Domain Data Augmentation

The purpose of this experiment was to find whether out-of-domain data could be used to improve the recognition rates of the original model. Table 2 shows that recognition rates for both splits improve from augmentation but do not benefit to the same extend. The German Split 1, which was poorly represented in its training data in ATCOSIM greatly improved. The already well represented Swiss split still actually improved, however only marginally. This shows that the usage of out-of-domain data can be particularly beneficial in cases where the test set is not well represented by the original corpus. However, Table 2 also shows that the retraining decreases the recognition rates for both of the splits with respect to the augmented model. This indicates that due to the small size of the corpus adding out-of-domain data improves the generalizing ability, which is then lost by retraining as it over-fits to the limited initial corpus. Additionally an interesting observation can be seen in the convergence of the of the retrained Swiss Split shown in Figure 2 when compared to the original Swiss Split convergence seen in Figure 1. First the convergence is much faster, which is to be expected. However, more interestingly is that the loss on the validation set is also lower than the validation loss of the non-retrained model, interestingly this does not translate into an actual improvement of recognition rates when comparing the WERs shown in Table 2.

C. Performance of Realistic Data

The purpose of this experiment was to investigate a model on more realistic data. In Table 3 it can be observed that the recognition rates of of the reference paper are much better than that of this research. The reference paper however mentions that Air and Ground are trained on 54 and 78 hours of training data respectively, whereas the results of this paper are only trained on 3.9 and 4.6 hours of data. In addition the lexicon provided was incomplete and therefore the paper might have made use of a more polished lexicon. Finally the reference paper uses a different acoustic modeling method. However, nothing conclusive can be said about the lower performance as it can be a combination of any of these. From Table 3 it can again be observed that the speed perturbations do not have any meaningful impact on the performance of the model. However an interesting observation is that the combined set performs considerable worse than both Air and Ground. As Combined is just a utterance combining Air and Ground fragments, one might expect the recognition rates to be somewhere in between the recognition rates of Air and Ground. However it is most likely caused by the cepstral mean and variance normalization of the utterances. As this happens for the entire utterance and probably due to the large difference in acoustic characteristics of each

of the signals neither of the signals performs well. In addition the Combined portion of the corpus is the smallest, and hence is most susceptible to large amount of variance within the corpus. However this does indicate that to achieve better performance it is important to separate Air and Ground.

D. Corpus Creation

All of the corpora presented in Section II relied on the access to data from institutional actors, or they created the data using simulated environments. Although *liveatc.net* is a source of openly available data, re-sharing such data can be problematic and therefore requires careful inspection of the law. The usage of two different corpora in this research has shown that it is important to consider the transcription method when creating such a corpus. It should first be identified that the problem of ASR in ATC can be separated into two problems. The first is the actual transcription of the utterance into words, and the second is the understanding of these words, such as the call-sign and commands. The ATCC corpus used a transcription method which had a more direct relationship to the to the actual understanding of the utterance, such as: *speedbird 3 5 A*, where the A is pronounced as *alpha*. This however creates ambiguity, as *A* can also be pronounced as the letter *A*. Given that ATC phraseology is specifically designed to minimize ambiguity it seems much more appropriate to use a word-to-word transcription method as used by ATCOSIM, with the added benefit that it is probably more easy to integrate with out-of-domain corpora.

VII. Conclusion

This paper has used two freely available Air Traffic Control (ATC) corpora and a freely out-of-domain corpus to create Automatic Speech Recognition (ASR) model for ATC using several different approaches. This process has shown that when considering creating a corpus, it is important to consider the specific application and try to minimize the amount of ambiguity that arises from transcription. Additionally this research has shown that there is no meaningful improvement to be gained from trying to increase training data by speed perturbations, and therefore is not an easy solution to the lack of in-domain data. However, when the used in-domain corpus is small, it does meaningfully improve recognition rates when augmenting the corpus with out-of-domain data, especially when the accent in the application slightly differs from the in-domain corpus. Finally it is important to separate air and ground utterances as it greatly improves recognition rates.

References

[1] Helmke, H., Rataj, J., Mühlhausen, T., Ohneiser, O., Her, H., Kleinert, M., Oualil, Y., and Schulder, M., "Assistant-based speech recognition for atm applications," *Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2015*, 2015.

[2] Šmídl, L., Švec, J., Tihelka, D., Matoušek, J., Romportl, J., and Ircing, P., "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development," *Language Resources and Evaluation*, Vol. 53, No. 3, 2019, pp. 449–464. doi: 10.1007/s10579-019-09449-5.

[3] Cordero, J. M., Rodríguez, N., De Pablo, J. M., and Dorado, M., "Automated speech recognition in controller communications applied to workload measurement," *SIDs 2013 - Proceedings of the SESAR Innovation Days*, , No. January, 2013.

[4] Chen, S., Kopald, H., Chong, R. S., Wei, Y. J., and Levonian, Z., "Read back error detection using automatic speech recognition," *12th USA/Europe Air Traffic Management R and D Seminar*, 2017.

[5] Chen, S., and Kopald, H., "The closed runway operation prevention device: Applying automatic speech recognition technology for aviation safety," *Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2015*, 2015.

[6] Benesty, J., Sondhi, M., and Huang, Y., *Springer Handbook of Speech Processing*, 2008.

[7] Dahl, G. E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, 2012, pp. 30–42. doi: 10.1109/TASL.2011.2134090.

[8] An, K., Xiang, H., and Ou, Z., "CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2020-Octob, 2020, pp. 566–570. doi: 10.21437/Interspeech.2020-2732.

[9] ICAO, *Manual of Radiotelephony*, 2007.

[10] Hofbauer, K., Petrik, S., and Hering, H., "The ATCOSIM corpus of non-prompted clean air traffic control speech," *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008, pp. 2147–2152.

[11] Zuluaga-gomez, J., Veselý, K., Blatt, A., Motlicek, P., Klakow, D., Tart, A., Szöke, I., Prasad, A., Sarfjoo, S., Kolˇ, P., Kocour, M., Cevenini, C., Choukri, K., Rigault, M., and Landis, F., "Automatic Call Sign Detection : Matching Air Surveillance Data with Air Traffic Spoken Communications †," 2020. doi: 10.3390/proceedings2020059014.

[12] Godfrey, J. J., "NIST - Air Traffic Control Complete," `https://catalog.ldc.upenn.edu/LDC94S14A`, 1994.

[13] Pigeon, S., Shen, W., Lawson, A., and Van Leeuwen, D. A., "Design and characterization of the non-native Military Air Traffic Communications database nn-MATC," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2, No. June, 2007, pp. 1373–1376.

[14] Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.-A., Clot, V., Gemello, R., Matassoni, M., and Maragos, P., "The HIWIRE database - digital source," `https://catalogue.elra.info/en-us/repository/browse/ELRA-S0293/`, 2008.

[15] Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.-A., Clot, V., Gemello, R., Matassoni, M., and Maragos, P., "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," 2007, pp. 3–6.

[16] Hofbauer, K., Petrik, S., and Hering, H., "The ATCOSIM corpus - digital source," `http://www.spsc.tugraz.at/ATCOSIM`, 2008.

[17] Šmídl, L., "Air Traffic Control Communication," , 2011. URL `http://hdl.handle.net/11858/00-097C-0000-0001-CCA1-0`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[18] Šmídl, L., Švec, J., Tihelka, D., Matoušek, J., Romportl, J., and Ircing, P., "Design and development of speech corpora for air traffic control training," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, pp. 2849–2853.

[19] Shore, T., Faubel, F., Helmke, H., and Klakow, D., "Knowledge-based word lattice rescoring in a dynamic context," *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, Vol. 2, No. September 2012, 2012, pp. 1082–1085.

[20] Oualil, Y., Schulder, M., Helmke, H., Schmidt, A., and Klakow, D., "Real-time integration of dynamic context information for improving automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2015-Janua, 2015, pp. 2107–2111.