BACHELOR THESIS

ARTIFICIAL INTELLIGENCE



|  |
| --- |
| Adaption, Analysis and Evaluation of Wav2Vec2 on Air Traffic Control Communication Data |

*First supervisor:*

dr. L.F.M. ten Bosch

Taal en communicatie

[louis.tenbosch@ru.nl](mailto:louis.tenbosch@ru.nl)

*Second supervisor:*

Dr. A. B. LAST NAME

Department

E-mail

*Author:*

Karan Chand

Student number: s1033357



28 August 2022

**Abstract**

Air Traffic Control (ATC) can benefit greatly from Automatic Speech Recognition (ASR), as communication between controllers and pilots contain essential callsigns and commands. Errors in communication can be disastrous and should be minimized. ASR decreases workload and improves aviation safety. ASR models designed for the ATC domain are limited, moreover, a distinct inadequacy of Wav2Vec2 architectures exist. State-of-the-art (SOTA) models have been studied insufficiently. This work therefore set out to create ASR models using SOTA Wav2Vec2 model architectures, fine-tune and evaluate them on robustness in the domain of ATC. Word Error Rate Reduction and Character Error Rate Reduction of ~95.5% and ~96.1% were achieved on the best performing XLS-R model. The effect additional training data and an in-domain language model have, are evaluated as well. It was found that when a maximum of 10% of training data was used, performance had a negative correlation to WERR and CERR. An in-domain language model has ~33% decrease on WER and ~18% decrease on CER when applied to the XLS-R models. These results indicate that a solid contribution to the field of ASR for ATC has been made, supplying additional Wav2Vec2 models in the process.

Contents

[1 Introduction 4](#_Toc112858170)

[2 Related work 6](#_Toc112858171)

[3 Methodology 7](#_Toc112858172)

[4 Experiments 11](#_Toc112858173)

[Data 11](#_Toc112858174)

[Settings 12](#_Toc112858175)

[5 Results and Discussion 12](#_Toc112858176)

[6 Conclusion 18](#_Toc112858177)

[7 Bibliography 19](#_Toc112858178)

[8 Appendix 22](#_Toc112858179)

# 1 Introduction

When travelling, safety of passengers in dependent on operators of the vehicle. For example, in trains and airplanes communication between operators and traffic control is crucial. Callsigns belonging to vehicles and command perception are essential for flawless execution and errors in communicating them should therefore be minimized. Focus on communication makes an already high workload task more difficult, which is why there is a need for assistance. To narrow down the topic, this paper focusses on Air Traffic Control (ATC). Communication between air traffic controllers and pilots is considered.   
ASR can be used to make ATC less complicated and error prone by transcribing the speech data. The addition of ASR has been shown to decrease workload by a factor of 3 in ATC communication (Helmke et al., 2016). The study calls for improvement of models, as they are crucial to error detection and prevention. Many variants of Automatic Speech Recognition (ASR) models exist, but there is an inadequacy in the domain of ATC. ATC benefits from ASR, as errors in verbal communication are common and can be very costly.

ATC communication grammar does not follow normal speech conventions. Utterances contain callsign and regularly commands, such as *‘lufthansa five three one eight contact zurich one three four decimal six’* and *‘swissair six five two zero rhein do you read’*. This is why ASR that has been trained on regular speech data performs poorly. In combination with noisy microphone audio, speech recognition becomes more complicated.  
In addition, data needed for training and evaluation of ASR models is scarce. Most transcribed ATC data is either publicly unavailable due to data privacy or of poor quality, due to noisy conditions, rate of speech and language accents. Manual transcription of publicly available audio data is costly and has been done before, however larger audio corpora are difficult to obtain (Zuluaga-Gomez et al., 2020).

In this research, multiple sequence-to-sequence models are considered. Sequence-to-sequence models are commonly used for ASR applications, as they predict data well, given future and past audio (Chiu et al., 2018). Artificial Neural Networks (ANN), Recurrent Neural Network (RNN) and Hidden Markov Models (HMM) have been used for ASR(Trentin & Gori, 2001). Only recently has the end-to-end Wav2Vec2 architecture been introduced and it is performing well (Baevski et al., 2020). The state-of-the-art (SOTA) transformer architecture has been shown to give a lower Word Error Rate (WER) than RNN and HMM based algorithms (Zuluaga-Gomez et al., 2022). Even online streaming of transformer models is developed, when it is generally limited to offline use (Moritz et al., 2020).  
ASR models used for ATC transcription do exist, but constant improvement is needed, as a perfect and generalisable model does not exist (yet). ASR models minimize errors most commonly based on WER (measures the percentage of words that are incorrectly predicted) and are massively in construction. SOTA ASR models should be adapted to ATC data to further minimize errors, as it can reduce risk of accidents and miscommunication (Wise et al., 1991).  
Models of ASR are designed and updated constantly, which leads to great innovation in this field. An example of a community that specializes in Wav2Vec2 transformer models and data sharing is HuggingFace (HF). Wav2Vec2 is a framework for self-supervised learning of speech representations, and has been shown to perform well on a broad spectrum of ASR tasks (Baevski et al., 2020). At HF, individuals can create, modify and publish their own models and datasets for others to use. Big companies such as Meta and Google have shared their models as well, which are trained on massive amounts of data. These models are very recent and have been tested on ATC data inadequately.

In this paper, Wav2Vec2 transformer models from HF are evaluated on their performance on transcribing ATC communication data from the simulated ATCOSIM dataset. The models that are considered contain:

* A model that is pretrained on relatively clean LibriSpeech data (Base).
* A model that is pretrained and fine-tuned on multiple noisy datasets as well as clean data (Robust).
* An XLS-R model, which is known for great multilingual and multi-accent performance and pretrained on the CommonVoice corpus.
* A huBERT model, which either improves or matches Wav2Vec2 performance and extends the Wav2Vec2 architecture. Also pretrained on LibriSpeech data.

This research means to fill the gap that is found in transcribing ATC data. Wav2Vec2 transformer models are evaluated on their performance and a road is paved for further research. The original models are evaluated on the ATCOSIM dataset - without adapting the models to the data - to give insight on how they perform. They are then adapted on the data to show their improvements. A language model is added to observe further decrease WER and CER. The contribution to the field is covered by the questions below.

**How robust are pretrained and fine-tuned ASR models on ATC data?**The main research question, as the answer gives us a broad explanation on the adaption, analysation and evaluation of models that are adapted on ATC data. Word Error Rate (WER) and Character Error Rate (CER) are considered metrics, while errors not explained well by them are evaluated.

The answers to the main research question are supported by two questions that are more direct and are answered by the experiments named above.

**How does a language model affect the performance of a fine-tuned XLS-R model on ATC data?**Performance enhancement caused by a domain specific language model can boost ASR models greatly (Oualil et al., 2015). General ASR models are pretrained on regular speech models, which do not include most ATC grammar. It is interesting to show the effect a specialized language model has on WER and CER.

**How does the amount of data used to fine-tune affect performance of a fine-tuned XLS-R model on ATC data?**Improving performance can be very cost-ineffective as great computational power is needed for training/fine-tuning a HF ASR model (Zuluaga-Gomez et al., 2022). Studying the amount of data needed for sufficient performance is valuable, as it aids and inspires further experiments.

In this paper, related work on the transcription on ATC data will be discussed. Then, reasoning and explanation of design choices are made. Experimental data is augmented and settings are shown. Next, performance of the models is reported, implications are made and questions are answered. Finally, research is summarized and concluded.

# 2 Related work

**Architectures**  
Read-back is the verbal confirmation of information by the receiver of communication. Reducing read-back errors in ATC has been researched before using an RNN and Long Short-Time Memory networks (JIA et al., 2018)(Lin et al., 2021). These models are able to get a relatively low WER (<10%), however, the models are still highly error prone. Wav2Vec2 is the most prominent model architecture, as it is high-performing and cost-effective towards training data (Baevski & Mohamed, 2020). Wav2Vec2 is flexible and can thus be fine-tuned on low resource datasets, such as Japanese and Arabic in the CALLHOME corpus (Yi et al., 2020). Yi et al. show that performance of Wav2Vec2 is universal over languages, though English does perform significantly better (52% improvement compared to ~25% in Japanese and Arabic). Model performance is improved by ~20% compared to previous work. Their research provides support for the research performed here, as ATC data can be compared to the datasets. ATC data is low resource and is different from regular English speech as well. Though ATC is fully spoken English in the ATCOSIM dataset, it follows a different grammar.

**Domain Specific Wav2Vec2**  
The domain of ATC has not been studied extensively, however, important works have been published and inspired this research. Domain shifting self-supervised learning of speech recognition is studied before (W. Hsu et al., 2021), however only one comparable study has been performed on ATC (Zuluaga-Gomez et al., 2022). Zuluaga-Gomez et al. published multiple papers on transcription of ATC data and callsign detection. In their paper, they discuss the feasibility of real-time ASR for ATC, study variable labelled data for fine-tuning and study the robustness of pretrained end-to-end models. The study does not evaluate types of errors made and is primarily focused on WER performance. Zuluaga-Gomez et al. manage to achieve an ~80% relative Word Error Rate Reduction (WERR) when training on all 14 hours of available speech data, which is promising, but it can be reduced further. In this paper, it is not the objective to reach a relatively low WER, as domain insight is the main objective.

**XLS-R Accents**XLS-R signifies cross-lingual speech representation. The definition of this architecture leads to the conclusion that it has an affinity for multi-lingual or accented speech data. This has been shown by the developers of the model (Babu et al., 2021) and by external parties as well (Kukk & Alumäe, 2022). Kukk and Alumäe have achieved a relative error rate decrease of 65% in a transcription task containing accented speech. They used the (at the time) SOTA XLS-R-300M model, which is pretrained on 500K hours of speech data from 128 languages. The research conducted surrounding XLS-R has been the main reasoning behind pursuing the XLS-R model for fine-tuning.

# 3 Methodology

**HuggingFace**  
Numerous python libraries exist for the development of ASR, such as scikit-learn and NLTK. These are standalone packages that are highly useful for tuning hyperparameters and developing NLP. For this research, complex development is not needed, rather testing of existing models is the focus. Testing of new and popular models can be more difficult to implement and test in these environments, as they are not easily accessible yet.   
The platform HuggingFace is more accessible. The HF community provides a wide range of publicly accessible pretrained models and datasets that can be modified with ease and reuploaded to the HF hub for public use. Over 25 transformer tasks are available for use, e.g. ASR, image classification and text generation. For clarification, if a model is pretrained, it has already been adapted on data before. If a model is fine-tuned, it has been trained on a downstream task, such as ATC data transcription. So, a model can both be pretrained and fine-tuned. In this paper, fine-tuning and training are used interchangeably. Models and datasets are constantly updated, which implies that SOTA tools are attainable. HF has a course for beginners and a decent documentation of the main libraries used: Datasets and Transformers. The learning curve is high, but the models and data delivered outweigh the costs compared to other libraries.

**Development**Different platforms are used for the development of transcribing, evaluating, training and testing of ASR models. HF recommends using Google Colab for ease of programming and the absence of needing to owning a high-end GPU for computation. For low computational power demanding tasks, such as small batch transcription, Google Colab is used, but the free to use platform has limitations that are diminishing our resources. GPUs can be either T4, P100 or a K80. These GPUs are inconsistently allocated by Google Colab and are unpredictable. Fine-tuning models and transcribing full datasets regularly exceed the resources (GPU ram, disk space and system ram) made available by Google Colab, leading to resource deprivation.   
For this reason, local development in scripts is done in addition to cloud development. A GeForce GTX 1080 GPU is used for transcribing full datasets and training of models for more robust development.

**Models**  
The models used are imported from the HF community. They are not trained or fine-tuned specifically on ATC data, as that domain has not been explored/published by the community yet. It is important that the models are recent, as ASR models are consistently updated and improved upon. Four models were considered for evaluating Wav2Vec2 models:

Base: facebook/wav2vec2-base-960h.   
Currently the most popular ASR model on HF that is updated approximately every 3 months. It has been pretrained and fine-tuned on 960 hours of Librispeech data. This model is called ‘base’ as it is mostly used for further training on a downstream task (finetuning). This model works well when clean, regular speech data is fed.

Robust: facebook/wav2vec2-large-robust-ft-swbd-300h.   
This model is a pretrained and fine-tuned version of the facebook/wav2vec2-large-robust model. The latest update to this model was 5 months ago. The data it has been fine-tuned on 300 hours of the Switchboard corpus. This is a telephone speech corpus that contains noisy data. This model has been chosen to see exactly how robust a robust model is when transcribing ATC data.

huBERT: facebook/hubert-large-ls960-ft.   
This model has been fine-tuned on 960 hours of Librispeech data and had its currently last update three months ago. The huBERT model follows the Wav2Vec2 architecture and extends upon it. huBERT’s training process is fundamentally different from Wav2Vec2, which the specifics of are out of scope. It has been shown that the model either improves or matches the performance of Wav2Vec2 (W. N. Hsu et al., 2021), so it is interesting to see if it performs better than the base and robust model on transcribing ATC data.

XLS-R: jonatasgrosman/wav2vec2-large-xlsr-53-english.  
The XLS-R model has been used for learning multilingual speech representation, dialect identification and accent identification (Pascal & Dominique, 2021). These challenges are apparent in ATC transcription as well, even though they are not the main objectives. The XLS-R model has been chosen, because the architecture uses contrastive learning and might perform better compared to the other models as a result. This specific model is updated regularly, with the last update pushed to HF being less than a month ago since beginning research.

**Evaluation**  
Evaluation of the model transcriptions was done by using the Word Error Rate (WER) and Character Error Rate (CER). These are the standard for ASR models as the accuracy from speech to text is the most important here (Malik et al., 2021). Metrics such as Glue or Bleu are not useful here, as they give information on the meaning of the output text compared to the reference text. These metrics are not needed in our ASR models and are thus omitted.  
The perplexity metric is used as well to determine how rare sentences are to appear in a gpt-2 text generation model. This model is trained on the text on 45 million Reddit web pages and contains 1.5 billion parameters. The gpt-2 model has been chosen, because it is the most popular text generation model found on HF with over 12 million downloads.

|  |  |
| --- | --- |
| **Training steps** | **Hours** |
| 10 | 0.011 |
| 50 | 0.053 |
| 150 | 0.158 |
| 500 | 0.528 |
| 1000 | 1.056 |

**Fine-tuning**   
Fine-tuning (training) is done using the standard template given by user ‘patrickvonplaten’ on the HF blog. Hyperparameters are not tuned to the dataset and models, as resources are limited. Training is done locally, as Google Colab has limitations that are mentioned above.   
The number of training data steps in which the models were fine-tuned in are chosen from the three defined distributions: low, medium and high data ranges. Ranges being from 1-100, 101- 500 and 501-inf respectively. Due to time and resource restrictions, only five models are fine-tuned.

* Two models in the low data range: 10 and 50 training samples.
* Two models in the medium data range: 150 and 500 training samples.
* One model in the high data range: 1000 training samples.

*Table 1: Ranges converted to hours*

**Addition** **Language** **model**  
The language model was made using an ARPA file containing only the training and validation transcriptions. Excluding the test cases from this file improves integrity and reduces overfitting. Overfitting is already a problem, as the dataset is quite small and only contains 743 unigrams. Because the word count is low, 1-5-grams are stored. Higher/lower N-grams are available, but are not studied here. Including a language model can boost the performance of ASR and is tested in this paper.

**XLS-R model**The reasoning behind choosing to finetune only the XLS-R model consists of three reasons. Firstly, the model is evaluated to be the most interesting to research on ATC data, as it has been shown that the model handles accents and multilingual data very well in comparison to the other models (Babu et al., 2021). Secondly, this model is exceedingly recent, as updates are posted regularly. It could have been interesting to fine-tune all the models; however, our resources would not grant this, producing the third reason. Given this information, the XLS-R model is the best option.

# 4 Experiments

## Data

There is a distinct shortage of publicly available data for ATC that can be used for ASR, as most datasets do not include transcriptions (Badrinath & Balakrishnan, 2022). Websites such as ‘*liveatc.net*’ provide, as the domain name implies, live and very recent ATC communication audio. Based on airport codes and frequencies, different radio communication is selected and listened to. This data does not include transcriptions and is therefore not suitable for training our models. Audio data is abundant and publicly available; however, audio data has to be manually transcribed, which is costly and time-consuming, leading to lower accessibility of transcribed ATC data (Zuluaga-Gomez et al., 2020).   
In this research, the ATCOSIM dataset is used. This corpus provides ten hours of speech data, pronounced by ten non-native speakers in the English language. A subset of the dataset can be found in the appendix. It includes additional information on the recording sessions and speakers, such as: Gender, speaker id, recording length, etc. This information was not used/ researched as it is out of the scope of this paper.  
The ATCOSIM dataset is a simulation of real-time ATC. The data is considered clean, as speech recordings were made in a controlled environment (Hofbauer et al., 2018)). Transcriptions have an average duration of 3.8 seconds. The same ATC lingo is used in the 10-hours of speech data. The data is noisy, as a close-talk headset microphone is used. Speaking style, language use, background noise and stress levels were kept as realistic as possible, but the representability of ATCOSIM to real ATC data has not been studied yet. Another limitation of this dataset is that it only contains speech of controllers and thus excluding pilots. Variables from the pilots’ side such as background noise are neglected, which should be taken into consideration when evaluating the generalisability of the studied ASR models.

The perplexity of the transcriptions of ATCOSIM is compared to the Google/fleurs corpus from HF. The Google/fleurs dataset contains sentences that are much more common in regular speech than the ATCOSIM dataset. The mean perplexity of the test cases from the Google/fleurs dataset is 35.98, while the ATCOSIM bears a staggering 175.08 mean perplexity. This shows that the ATCOSIM significantly diverges from normal speech and constitutes to low performance of regular speech models on this data.

The ATCOSIM dataset can be downloaded from the internet and can be obtained on DVD. It is not yet available on HF. The corpus is formatted in ISO, which needs to be reformatted for HF transformers use. As mentioned above, additional information on speakers and recordings are removed from the dataset and will not be used.  
Additionally, unusable rows in the dataset are removed. Unusable meaning:

* Rows containing transcriptions that exceed the maximal duration allowed for training and model transcription, as GPU usage is a limitation in this research.
* Rows containing information that is chosen not to be included for research as it is either out of the scope or to increase efficiency of ASR. Such information includes: Human noise, word fragments, empty utterances, off-talk, nonsensical words, foreign language and unknown words. These are represented by the tags respectively: [HNOISE], [FRAGMENT], [EMPTY], (<OT> … </OT>), [NONSENSE] and (<FL> … </FL>). Where three dots represent the speech that corresponds to the tags.
* The characters in the transcription that are regarded as not useful for training and transcribing. Keeping the transcriptions as simple as possible generally increases efficiency of ASR models. Characters that were removed, separated by ‘*\*’: [*\=\~\@\,\?\.\!\-\;\:\"\“\%\‘\”\�\'*].

The ATCOSIM corpus originally has 10078 rows, however after cleaning up and pruning, 9397 rows remain. The train-validation-test split is made using an 8:1:1 ratio, which correlates to 7526-940-931 rows respectively. However, due to resource limitations, only a maximum of 1000 training rows and a constant 50 validation rows are used for training. The number of test cases stay consistently 931.

## Settings

All training experiments use the same set of hyperparameters. The feature encoder is frozen during the fine-tuning phase, as it is trained sufficiently during pretraining. We fine-tune each model for 7.5k steps, with a 100-step warm-up phase. We fine-tune each model on a GeForce GTX 1080 with an effective batch size of 8 (batch size of 4, gradient accumulation of 2). Learning rate is increased linearly until 3e−4 during warm-up, then it linearly decays.

# 5 Results and Discussion

First, the pretrained models that are directly pulled from HF are evaluated. In figure 1 it shown that the WER and CER are relatively high. These models were not given an explicit language model. The base model performs the worst, which is expected, as it has not been pretrained on noisy speech data. It has the lowest WER and CER. The base model is the oldest model that is researched, which should be noted.   
The robust model performs better compared to the base model, as it is pretrained and fine-tuned on noisy telephone speech data. This results in a significant reduction in WER and a decent reduction in CER.  
The huBERT model possesses the smallest WER and CER. Improved or identical performance compared to standard Wav2vec2 models is expected from the huBERT model (W. N. Hsu et al., 2021). A modified architecture might be the cause of this, as the model has been fine-tuned on the same data as the base and robust model.  
The XLS-R model was expected to have a lower WER than the base and robust model, as it ought to have an advantage in transcribing data that has multiple accents (Zuluaga-Gomez et al., 2022). However, this is not the case. The cause for this might be that the robust model has a bigger advantage, as it is pretrained on noisy data, in contrast to the XLS-R model.  
The CER is relatively low for all models, which indicates that the models are able to recognize letters being spoken, however, they do not have an in-domain language model. Airport names and callsigns are uncommon in regular speech, relating to a high WER.

Figure 1: Evaluation of pretrained models in WER and CER.

To continue, the XLS-R model is fine-tuned on increasing amounts of data (figure 2). We see a massive drop in WER and CER as the amount of data increases. The WERR for the XLS-R-10 model is already at ~39%. The highest WERR is seen found in the XLS-R-1000 model at ~93%. CER Reduction CERR for the XLS-R-1000 model is significant as well (~92%).

Figure 2: Evaluation of the fine-tuned XLS-R models in WER and CER.

Next, a language model is added to the Wav2Vec2 processor. Absolute reduction of performance seems to be variable, but relative reduction of the WER and CER per model achieves ~33% decrease on WER and ~18% decrease on CER. The model with the lowest WER and CER is, as expected, the XLS-R-1000\* model. However, the WERR of the lower data provided models are significant and impressive as well.

Figure 3: Evaluation of the fine-tuned XLS-R models in addition to a language model in WER and CER.

|  |  |  |
| --- | --- | --- |
| **Model** | **WERR** | **CERR** |
| xls-r-10\* | -32% | -21% |
| xls-r-50\* | -44% | -29% |
| xls-r-150\* | -30% | -17% |
| xls-r-500\* | -33% | -18% |
| xls-r-1000\* | -33% | -16% |

Table 2 WERR and CERR of the models including an in-domain language model, in comparison to the models without the language model

The overall improvement of the XLS-R model is shown in figure 4. It shows that fine-tuning on ATC data can significantly improve the overall performance on WER and CER, especially in addition to a language model.

Figure 4: Improvement of the XLS-R model in WER and CER

**Comparison to Zuluaga-Gomez et al. 2022**  
Similar work on fine-tuning older models has been done on ATC data (Zuluaga-Gomez et al., 2022). Zuluaga-Gomez and colleagues’ lowest utilization of training data consists of 5 minutes of speech and achieved WERs of 40% and 43.9% for ISAVIA and NATS respectively. Air navigation service providers collected and annotated these datasets, which originated from actual ATC communication. As a different dataset is used for evaluation, a comparison between the models does not seem logical. Especially since the data that is used here is not actual ATC data. However, it should be mentioned that the WER and CER acquired from the XLS-R-1000\* model are extremely low, considering only 1 hour of training data was used to fine-tune the model.  
The lowest utilization of training data in our research is 38 seconds and achieved a WER of 43% without an addition language model. Addition of a domain specific language model further decreases WER to 0.29, resulting in a better ASR model using less training data. Again, these model performances should not be compared to the work that has been done by Zuluaga-Gomez and colleagues, rather it should give insight on the model capability.

**Unusually high performance - Overfitting**  
The definite WERR and CERR of the XLS-R-1000\* model compared to the original pulled xls-r model from HF amount to ~95.5% and ~96.1% respectively. Such high rates are unusual and could be explained by the origins of the ATCOSIM dataset. As it is simulated, data is similar. Ten speakers are used to fill the corpus and recordings have the same audio quality.  
Also, all speakers were integrated in the train, test and validation set. Even though the model is tested on unseen data, the model is most likely overfitting the data. Therefore, it would have been better to isolate speaker(s) from the train and test set during preprocessing.  
Another reason for high performance is overfitting of the language model. Since the model was constructed on 90% of available ATCOSIM data. The remaining 10% of the test data is too similar to training and validation set to label these models as generalizable.

**Evaluation XLS-R-1000\***  
The best performing model of this research has shown to be the XLS-R-1000\* model. Examination of errors made by this model is valuable, as it gives insight on the usability of the model in practice.

First, it has been shown that WER and CER are relatively low (<5%) compared to the standard models pulled from HF (figure 4). However, callsign detection is essential in the ATC domain. A common error made by the model is found at the start of the callsign, as the model wrongly predicts the names of the airline codes. Predictions made are still recognizable, but they are wrong. For example, *‘golf’* is predicted to be *‘goll’*, which is useful information, as currently no other airline code sound or looks close to this word. Still, it is wrongly predicted and contributes to the performance drop of this model.   
The model is able to recognize numbers well, as the model rarely predicts them wrong. For example, *‘two’* is predicted as *‘to’*, but it is discernible to recognize as the number ‘*two’*. Even though these errors are rare, they are present.

Second, speech files containing only a single word are regularly incorrectly predicted (appendix table 2). The reason for this is that the model highly depends on the language model when it is not sure which word it should predict based on audio quality alone. However, the language model has no context before or after the singular word is shown. This leads to the model basing its prediction on noisy audio quality, which is highly error prone.   
The impact of these errors on practical use of this model is negligible. Errors in single word communication is not the main cause of ATC miscommunication (Helmke et al., 2016). Complete and flawless callsign delivery is of more importance, as the essence of the message is encapsulated in it.

Third, the model frequently adds/subtracts non-lexical conversation sounds, such as: *‘ah’* and *‘oh’* (appendix table 2). These small errors are not considered to be detrimental to performance, as the manually transcribed data is not consistent. Having listened to the audio data that corresponds to these transcriptions, it is found that the presence of these sounds can be ambiguous. These errors should therefore not be detrimental to performance, contradictory to WER and CER. The addition or subtraction of these words should have no shift in meaning of the message. Is a different metric more suitable?

**Evaluation Metric**  
It has been shown above that the model makes different errors, but where does the focus lie? Errors such as *‘ah’* and *‘oh’* does not change the concept of the information sent/received. The essence of information is not captured by WER or CER, just the crude errors words and characters.  
Should a different metric than WER and CER be used? In a study conducted by Oualil and colleagues two additional ATC specific metrics are used to more accurately represent performance of transcribing ATC data (Oualil et al., 2015). Concept Error Rate (ConER) and Command Error Rate (CmdER) are introduced. ConER makes use of labelled transcribed data to determine if the utterance is either a command or a callsign. CmdER requires all concepts in the utterance to be correct. WER and CER can be compared to CmdER and ConER respectively, as CmdER determines performance on a broad level (macro – word based) and ConER determines performance in a narrower view (micro – character based).   
Specific ATC metrics would have been interesting to research given the ATCOSIM corpus and the fine-tuned models, as WER and CER do not fully capture performance (Srinivasamurthy et al., 2017).

**Further research**Models on HF have not been studied intensively on ATC data, which leads to many different opportunities for further researching the field. Questions that could not be answered during this research due to breach of focus or resource constraints are documented here to inspire further research.

Only one recent pretrained model is fine-tuned in this paper. To fully capture the performance Wav2Vec2.0 models have on transcribing ATC data, other (mentioned) models can be researched as well. This has been done before (Zuluaga-Gomez et al., 2022), but SOTA ASR models are developed commonly on a weekly basis and should be examined on their performance when the domain is shifted.   
Fine-tuning on variable amounts speech data to show the effect it has on performance. This has been done in this paper; however, fine-tuning was limited to only ~1 hour of speech data due to GPU restrictions. Deploying the experiments on a higher scale would certainly give more insight on efficient data usage.   
Hyperparameters were not tuned during training. Standard procedures were followed from the HF community and were sufficient for conducted research. Tuning the training settings leaves possibilities for further improvement of the models, training efficiency and conceptual comprehension of training on ATC data.  
Publicly available transcribed ATC audio is scarce and a limitation of research, hence performance generalisation of the models is limited. Evaluation on various (real) ATC audio data is required for further development of the models.  
Performance of variable language models can be researched as well. Only a single language model is used here, as focus is kept on the improvement *a* language model has on the task at hand. For further research, multiple language models should be tested and evaluated on performance increase. Language models should include variable N-gram sizes and in/out data origin construction for completeness. The language model considered here originates from the same dataset that is used for training, which can lead to overfitting, but needs confirmation.  
A finite state machine can be used for creating a language model (more) independent of training data (Oualil et al., 2015). Huge amounts of text data can be constructed and compiled to generate a language model. The finite state machine provides a blueprint for callsigns and commands, which can decrease errors shown above. Numerical callsign errors such as *‘two’* predicted as *‘to’* should be less common, as the grammar increases the probability that the ASR model predicts *‘two’* when the word is situated in a callsign (Nguyen & Holone, 2015). In practice, all airline names are known to air traffic controllers, so names can be added to the language model for completeness and might increase model accuracy.

**How robust are pretrained and fine-tuned ASR models on ATC data?**  
The models that were imported directly from the HF community show very low robustness without any fine-tuning. Lowest achieved WER and CER are 51% and 22% respectively, which are attained by the huBERT model. However, after fine-tuning the XLS-R model without an in-domain language model, the WER and CER already drastically drop, with the best model (XLS-R-1000) already achieving a reasonable WER of ~5% (Helmke et al., 2016). Addition of a language model further decreases WER and CER, achieving a high performance XLS-R model. Robustness of the models directly pulled from the HF is thus very low on ATC data, though minor training and an in-domain language model can have a major impact.

**How does a language model affect the performance of a fine-tuned XLS-R model on ATC data?**It has been shown above that the in-domain language model has ~33% decrease on WER and ~18% decrease on CER on XLS-R models. The language model gives a performance boost in exchange for low additional computation power. For example, it can be deduced that *‘to’* should be interpreted as *‘two’* when the word is contained in a callsign. This can be easily implemented using a language model, while it might take more data/training to receive the same result.

**How does the amount of data used to fine-tune affect performance of a fine-tuned XLS-R model on ATC data?**In figure 2 the models without a language model are shown in ascending order of training data used. There does not seem to be a performance decrease when fine-tuning on 1000 rows of data, which leads to the conclusion that it might be possible to increase the amount of training data further, potentially achieving higher performance. Overfitting might be even more apparent than shown above, though this needs further work for confirmation. In this paper however, it seems clear that more training data equals higher performance, given that we only use a maximum of 10% of training data.

# 6 Conclusion

This paper adapted, analysed and evaluated pretrained and finetuned Wav2Vec2 models on ATC transcription using simulated ATC communication data. Models were compared using WER and CER, with the best model (XLS-R-1000\*) achieving ~3% and ~1.5% respectively. WERR and CERR of ~95.5% and ~96.1% respectively were reached on the ATCOSIM dataset. Errors that were still made by the best model contain callsign detection, single word prediction and addition/subtraction of non-lexical words. Callsigns are still recognizable, but perform poorly according to WER and CER; other metrics such as CmdER and ConER should be used. Furthermore, it was demonstrated that models directly imported from HF are not robust considering ATC data, though minor training will greatly improve performance. Next, an in-domain language model extended the XLS-R models and it is shown that the language model has a significant impact on WER and CER on the lower data trained examples (high WERR and CERR), though the effect decreases when increasing training data. Lastly, variable amounts of training data were plotted against WER and CER to show that there is a negative correlation between them, which did not change when 10% of training data was used.

# 7 Bibliography

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2021). *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. http://arxiv.org/abs/2111.09296

Badrinath, S., & Balakrishnan, H. (2022). Automatic Speech Recognition for Air Traffic Control Communications. In *Transportation Research Record* (Vol. 2676, Issue 1, pp. 798–810). SAGE Publications Ltd. https://doi.org/10.1177/03611981211036359

Baevski, A., & Mohamed, A. (2020). *Effectiveness of Self-Supervised pre-training for ASR*. IEEE.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. http://arxiv.org/abs/2006.11477

Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 4774–4778. https://doi.org/10.1109/ICASSP.2018.8462105

Helmke, H., Ohneiser, O., Muhlhausen, T., & Wies, M. (2016). Reducing controller workload with automatic speech recognition. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2016-December*. https://doi.org/10.1109/DASC.2016.7778024

Hofbauer, K., Petrik, S., & Hering, H. (n.d.). *The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech*.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). *Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training*. http://arxiv.org/abs/2104.01027

JIA, G., CHENG, F., YANG, J., & LI, D. (2018). Intelligent checking model of Chinese radiotelephony read-backs in civil aviation air traffic control. *Chinese Journal of Aeronautics*, *31*(12), 2280–2289. https://doi.org/10.1016/j.cja.2018.10.001

Kukk, K., & Alumäe, T. (2022). *Improving Language Identification of Accented Speech*. http://arxiv.org/abs/2203.16972

Lin, Y., Guo, D., Zhang, J., Chen, Z., & Yang, B. (2021). A Unified Framework for Multilingual Speech Recognition in Air Traffic Control Systems. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3608–3620. https://doi.org/10.1109/TNNLS.2020.3015830

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

Moritz, N., Hori, T., & le Roux, J. (2020). *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Nguyen, V. N., & Holone, H. (2015). *Possibilities, Challenges and the State of the Art of Automatic Speech Recognition in Air Traffic Control*.

Oualil, Y., Schulder, M., Helmke, H., Schmidt, A., & Klakow, D. (2015). *Real-Time Integration of Dynamic Context Information for Improving Automatic Speech Recognition*.

Pascal, F., & Dominique, R. (2021). *Bachelorarbeit (Informatik) Speech Classification using wav2vec 2.0*.

Srinivasamurthy, A., Motlicek, P., Himawan, I., Szaszák, G., Oualil, Y., & Helmke, H. (2017). *Semi-supervised Learning with Semantic Knowledge Extraction for Improved Speech Recognition in Air Traffic Control*. http://www.aclistant.de

Trentin, E., & Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. In *Neurocomputing* (Vol. 37).

Wise, J. A., Hopkin, V. D., & Smith, M. L. (Eds.). (1991). *Automation and Systems Issues in Air Traffic Control*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-76556-8

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). *Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages*. http://arxiv.org/abs/2012.12121

Zuluaga-Gomez, J., Prasad, A., Nigmatulina, I., Sarfjoo, S., Motlicek, P., Kleinert, M., Helmke, H., Ohneiser, O., & Zhan, Q. (2022). *How Does Pre-trained Wav2Vec2.0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications*. http://arxiv.org/abs/2203.16822

Zuluaga-Gomez, J., Veselý, K., Blatt, A., Motlicek, P., Klakow, D., Tart, A., Szöke, I., Prasad, A., Sarfjoo, S., Kolčárek, P., Kocour, M., Černocký, H., Cevenini, C., Choukri, K., Rigault, M., & Landis, F. (2020). *Automatic Call Sign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications*. 14. https://doi.org/10.3390/proceedings2020059014

# 8 Appendix

Full list of code, data and output files can be found at <https://github.com/KaranChand/ATC>. This includes data augmenting, fine-tuning and evaluation of ATC data.  
Augmented data can be found on HuggingFace: <https://huggingface.co/KaranChand>  
All models and data used are uploaded and available for personal use.

Table 3: A subset of the ATCOSIM dataset:  
  
More information on the dataset can be found at: <https://www.spsc.tugraz.at/databases-and-tools/atcosim-air-traffic-control-simulation-speech-corpus.html>

Table 4: A subset of the errors made by the (best performing) XLS-R-1000\* model:

|  |  |  |  |
| --- | --- | --- | --- |
| **transcription** | **recording\_id** | **model\_transcription** | **difference** |
| golf bravo victor juliett india is identified good afternoon | 051\_0227 | goll bravo victor juliett india is identified good afternoon | {'golf'} |
| exact | 111\_0627 | expect | {'exact'} |
| roger what is your position | 101\_0199 | roger ah what is your position | set() |
| japan air four one nine contact milan one three four five two bye | 101\_0308 | german air four one nine contact milan one three four five two bye | {'japan'} |

Difference represents the words that were not found in transcription, but were found in model\_transcription. That is why set() is found in the second to last row, as the model predicts extra words.