BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Automatic Transcription of Air-Traffic Communication to Text

Veronika Nevařilová*

Abstract

This paper serves as a brief description of the process of fine tuning Whisper, an automatic speech recognition model, on Czech and English air-traffic communication recordings. It proposes two different forms of transcription – typical, full transcriptions of the recordings, and abbreviated, for faster orientation in transcriptions. The model was fine tuned parallelly on both forms to see its difference in performance. Even though the training dataset was quite little, training was able to decrease word error rate (WER) of Whisper for full transcriptions on Czech data more than 6 times and on English data almost 9 times, while Whisper for abbreviated transcriptions significantly improved as well. This means that with larger datasets, Whisper may be likely to continue improving at learning new patterns even in recordings with large noise levels and possibly serve as a help for people working with air-traffic communication.

*xnevar00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Despite efforts to minimize errors in air-traffic communication through the use of standardized phrases and special pronunciation of certain words, it is not uncommon for air traffic controllers (ATCos) to mishear words or fail to understand them at all. This can easily happen primarily when most of ATCos handle more than one frequency/phone and multiple parties initiate transmission at the same time. Such situations can excessively delay the flow of air traffic controlling or even have serious consequences. To ensure greater certainty for air traffic controllers that they have understood correctly, having a textual transcription of the pilot's broadcast in front of them could make air traffic management safer and faster.

The goal of this work was to train a speech recognition model tailored for Czech-English air traffic communication potentially usable by ATCos or airport staff, e.g. for transcription of air-traffic communication recordings that are obligatory to be recorded at every airport.

2. Transcription protocol

Since air-traffic communication uses abbreviations and numbers a lot, it could be quite difficult for ATCos to navigate through fully transcribed text when searching quickly for some important information. For this purpose, two transcription protocols were designed.

- Full transcription – every word exactly as said:

  Oscar Kilo Alpha Bravo Charlie dráha nula dva střední přistání povoleno vítr nula jedna nula stupňů pět uzlů

- Shortened transcription – mainly numbers, call signs and other often used key information are abbreviated:

  OKABC dráha 02C přistání povoleno vítr 010 stupňů 5 uzlů

The reason for two transcription protocols was to train the model parallelly on full and shortened transcriptions and analyze its performance on standard full text transcribing and on shortened text, which is full of abbreviations, that the model has certainly never come across yet.

3. Used dataset

Audio data consist of Czech and English (and some minority of Slovak) recordings of Kunovice airspace (LKKU). The ratio of Czech:English recordings is

cca 80:20.

The recordings were all transcribed with a help of SpokenData[1]. They were transcribed using both types of transcription and then for each type of transcription protocol separate datasets were made.

Final training datasets contain over 1500 air-traffic communication recordings making it cca 5 hours of data in total.

4. Training

Training was realized by fine tuning OpenAI Whisper[2] Medium. It is an *encoder-decoder transformer* [1] multilignual model that has been pre-trained on 192 hours of Czech language audio [2]. Since it is not trained on any specific dataset, it is very flexible and thus suitable for fine tuning for specific usage.

Training script utilizes the Hugging Face Transformers[3] library, with which it is possible to train numerous models for speech recognition, computer vision etc., and the training itself took place on university's computing cluster.

Whisper was making quite large differences in performace among multiple trainings with same hyperparameters and dataset, but for example setting that achieved the best training performance on model for full transcriptions was as follows:

```
per_device_train_batch_size=1

gradient_accumulation_steps=16

learning_rate=1e-3

warmup_ratio=0.12

num_train_epochs=45
```

5. Results

As seen in Figure 1, the baseline model's word error rate (WER) on full transcriptions achieves 90.5 % on Czech and 143.4 % on English data when tested on LKKU test set. After training, both values have significantly decreased reaching 14.3 % on Czech and 16.3 % on English data.

Figure 2 shows the same information, but for the model for shortened transcriptions. WER on Czech data decreased from 104.2 % to 21.5 % after training. WER on English data for baseline model was 187.1 %, while after training it reached 27.5 %.

[1] https://www.spokendata.com/

[2] https://openai.com/research/whisper

[3] https://huggingface.co/docs/transformers/index

6. Conclusions

Whisper proved itself to be a flexible automatic speech recognition model for challenging tasks such as learning on recordings with significant noise level and memorizing many abbreviations of multiple words. 14.3 % WER on full transcriptions of Czech recordings makes the model already reliable when it comes to transcribing air-traffic communication. 21.5 % WER on shortened transcriptions is a sign that it can easily handle more complex tasks with as little as 5 hours of training data.

When experimenting with training hyperparameters, it was quite hard to find out what settings make big difference, since Whisper makes a little difference each training itself. This feature can also help with model's accuracy in a way that the model can sometimes overcome some local optima though.

I believe that with larger amount of training data, it would be possible to continue improving the model and to possibly create a system that could be used by airport staff to improve their efficiency.

Acknowledgements

I'd like to thank my supervisor Ing. Igor Szőke, Ph.D for their help, advice and for providing me with an opportunity to transcribe my recordings on SpokenData, which saved me a lot of time. I also very much appreciate arranging access to a computing cluster, since the training wouldn't be possible without it.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and et al. Attention is all you need. online, Jun 2017. https://arxiv.org/abs/1706.03762.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. online, Dec 2022. https://arxiv.org/abs/2212.04356.