Sequence-to-Sequence Speech Recognition for Air Traffic Control Communication

Tijs Rozenbroek
Thesis supervisors: Dr F.A. Grootjen and Dr U. Güçlü

Radboud University, Nijmegen, the Netherlands
t.rozenbroek@student.ru.nl

Keywords: Automatic Speech Recognition · Air Traffic Control · Sequence-to-Sequence Models

1 Introduction

Air Traffic Control (ATC) is absolutely essential in aviation, and air traffic controllers have the highly taxing job of directing all air traffic within a specific region, preventing accidents, giving information to pilots and more. Since ATC is so important, all efforts must be made to ensure that ATC communication happens optimally and efficiently, without error. Any effort to assist the air traffic controllers or pilots in their communication is therefore warranted.

An example of a system that can assist air traffic controllers and pilots is a system which detects errors in communication, such as errors in repeating instructions and callsigns, known as readback errors. What follows is an example of how an undetected readback error can lead to dangerous situations.

On the 7th of March 2016, at EuroAirport Basel Mulhouse Freiburg, a serious incident took place due to a combination of factors [1]. A readback error by a pilot was not corrected by the air traffic controller, which led to two planes being on the same runway at the same time. The planes missed each other by a mere 115 meters, which is a very small distance in these kinds of situations.

A system that could automatically detect this type of error, and warn pilots or controllers, could improve overall aviation safety. To build such a system, or a related system, a sufficiently fast and reliable automatic speech recognition (ASR) system is required.

Currently, few ASR systems have been developed for ATC on the whole. According to Helmke et al. [3], efforts to bring ASR into the domain of ATC have been made as early as the 1990s. However, no instances of using sequence-to-sequence models for ATC have been found, which is the gap in the field that this work aims to fill.

2 Background

2.1 Air Traffic Control

There are, unfortunately, factors that make the ATC domain a difficult domain to implement ASR into. These specifics are, amongst others, high levels of noise, non-native speakers (accents), standardised special phraseology and perhaps more importantly, deviations from this standard phraseology. Additionally, ATC communication is rapid and thus lightweight and fast ASR solutions are required.

2.2 Sequence-to-Sequence Model

The sequence-to-sequence model architecture that was taken as the basis for the experiments in this work, is the recently published model architecture by Hannun et al. [2]. The authors show that the model architecture performed well when trained and tested on the LibriSpeech dataset [4], where it attained a word error rate (WER) of 15.64 without an external language model, 11.87 in combination with a 4-gram language model and 9.84 with a convolutional language model, when evaluated on the slightly more challenging 'test-other' test set from LibriSpeech [2].

The model architecture's most interesting and novel feature is the use of time-depth separable (TDS) convolutions, which, as the authors claim, generalise better than other deep convolutional architectures and use fewer parameters.

3 Methods, Results and Conclusion

As mentioned, the model architecture by Hannun et al. was taken as the basis for the experiments in this work. Several approaches were taken for attempted improvements of the model architecture, ranging from increasing receptive fields of the aforementioned TDS convolutions, to increasing the amount of TDS layers. The training configuration was manipulated in several ways to improve convergence and thus improve performance.

The best-performing model that was made in this work, scored a word error rate of 26.19% on noisy, low-quality ATC data, and 5.9% on relatively clean data. It is important to mention that these tests were conducted without external language models, leaving room for further improvements. The high WER on the noisy data can be largely attributed to its noise, which caused some utterances to be nearly unintelligible, even to a trained ear. Addressing these issues would be key for improving performance, which could perhaps partially be done by improving the robustness of the model.

With these results in mind, it can be stated that in the future, sequence-to-sequence models in general might be a viable option for an ASR model for ATC, and time spent further developing these models would be well spent. All in all, a solid contribution to the field of automatic speech recognition for air traffic control has been made, since the absence of sequence-to-sequence models in this field has been concluded.

References

1. Serious incident to a dornier 328 registered HB-AEO and to an embraer 190 registered PH-EXB occured on 07/03/2016 at bâle-mulhouse (68) (2018), https://www.bea.aero/uploads/tx_elydbrapports/BEA2016-0122.en.pdf
2. Hannun, A., Lee, A., Xu, Q., Collobert, R.: Sequence-to-sequence speech recognition with time-depth separable convolutions. In: Interspeech 2019. pp. 3785–3789. ISCA (2019). https://doi.org/10.21437/Interspeech.2019-2460
3. Helmke, H., Ehr, H., Kleinert, M., Faubel, F., Klakow, D.: Increased acceptance of controller assistance by automatic speech recognition. In: Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013). pp. 1–10 (2013), https://elib.dlr.de/87600/
4. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210. IEEE (2015). https://doi.org/10.1109/ICASSP.2015.7178964