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

Tijs Rozenbroek
linkedin.com/in/tijs-rozenbroek
t.rozenbroek@student.ru.nl

Supervisor:
Dr F.A. Grootjen
Radboud University
AI department

Second reader:
Dr U. Güçlü
Donders Centre for Cognition
Donders Institute for Brain, Cognition and Behaviour
Radboud University AI department

1. Introduction

Air Traffic Control (ATC) is essential for aviation, and air traffic controllers have the highly taxing job of directing all air traffic within a specific region, preventing accidents and more.

Automatically detecting errors in ATC communication could improve aviation safety. To do this, a good automatic speech recognition (ASR) system is required.

Currently, few ASR systems have been developed for ATC. Sequence-to-sequence (s2s) models have not been used for ATC, which is the gap in the field that I aimed to fill.

2. Background

The ATC domain is difficult to implement ASR into. This is because of e.g. high levels of noise and non-native speakers (accents).

The s2s architecture that is the basis for my experiments, is the recent model architecture by Hannun et al. [1]. Its most novel feature is time-depth separable (TDS) convolutions (see Fig. 1), which generalise better than other architectures and use fewer parameters [1].

Two ATC corpora were used as data:

- **ATCOSIM corpus** [2] - 10 hours of quite clean data from real-time simulations, only controllers' speech.
- **ATCC corpus** [3] - 20 hours of noisy low-quality data from operational ATC, including pilots.

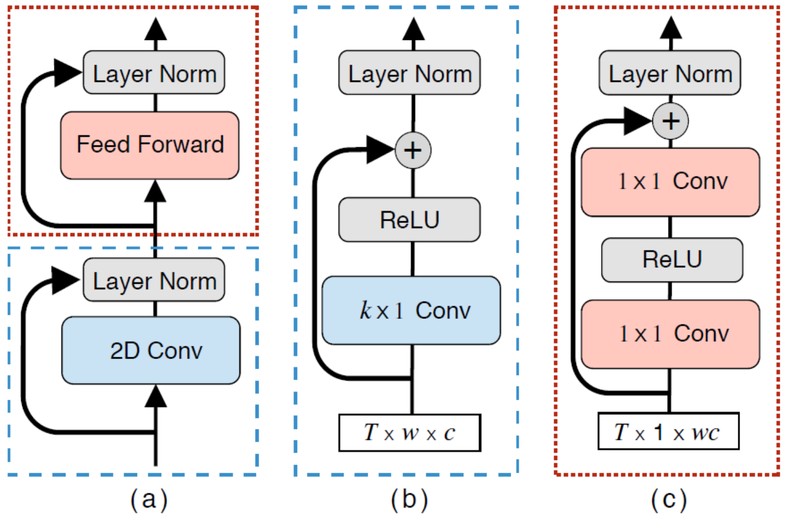

Figure 1: *"The TDS convolution model architecture. (a) The sub-blocks of the TDS convolution layer are (b) a 2D convolution over time followed by (c) a fully connected block."* Taken from [1].

3. Methods

The architecture by Hannun et al. is the basis of my experiments. Several approaches were taken to improve the architecture, for example:

- Bigger receptive fields of TDS convolutions.
- More TDS layers.
- Altering training config for better convergence.

4. Results

	WER	
Experiment	**ATCOSIM-test**	**ATCC-test**
Base	7.64	31.46
Improved	5.90	26.19

The base model performed well on the clean ATCOSIM test set, but poorly on the noisy ATCC test set. After improvements, the word error rates substantially improved.

5. Discussion

The WER of **26.19** on the ATCC test is still quite high. This can be mostly explained by the following notions:

- Noise
- Differing volumes
- Non-native speakers

The first two issues stem from the fact that ATC communication takes place over radio, where many different radio and microphone equipment types are in use, and their settings can differ.

No language models were used while testing, so improvements can be made there. Another improvement would be to make the models more robust to noise.

6. Conclusion

By using and improving the s2s architecture by Hannun et al. for ATC, the gap in the field of ASR for ATC, caused by the absence of s2s models, has begun to be remedied and a general contribution to the field has been made.

References

[1] A. Hannun, A. Lee, Q. Xu, and R. Collobert, 'Sequence-to-Sequence Speech Recognition with Time-Depth Separable Convolutions', in Interspeech 2019, Sep. 2019, pp. 3785–3789, doi: 10.21437/Interspeech.2019-2460.
[2] L. Šmídl and P. Ircing, 'Air Traffic Control Communication (ATCC) Speech Corpus', presented at the CLARIN Annual Conference 2014 - CAC2014, Soesterberg, The Netherlands, 2014.
[3] K. Hofbauer, S. Petrik, and H. Hering, 'The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech', in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May 2008, [Online].