The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR

Aiden Williams
Supervisor: Dr Andrea DeMarco,
Co-Supervisor: Dr Claudia Borg

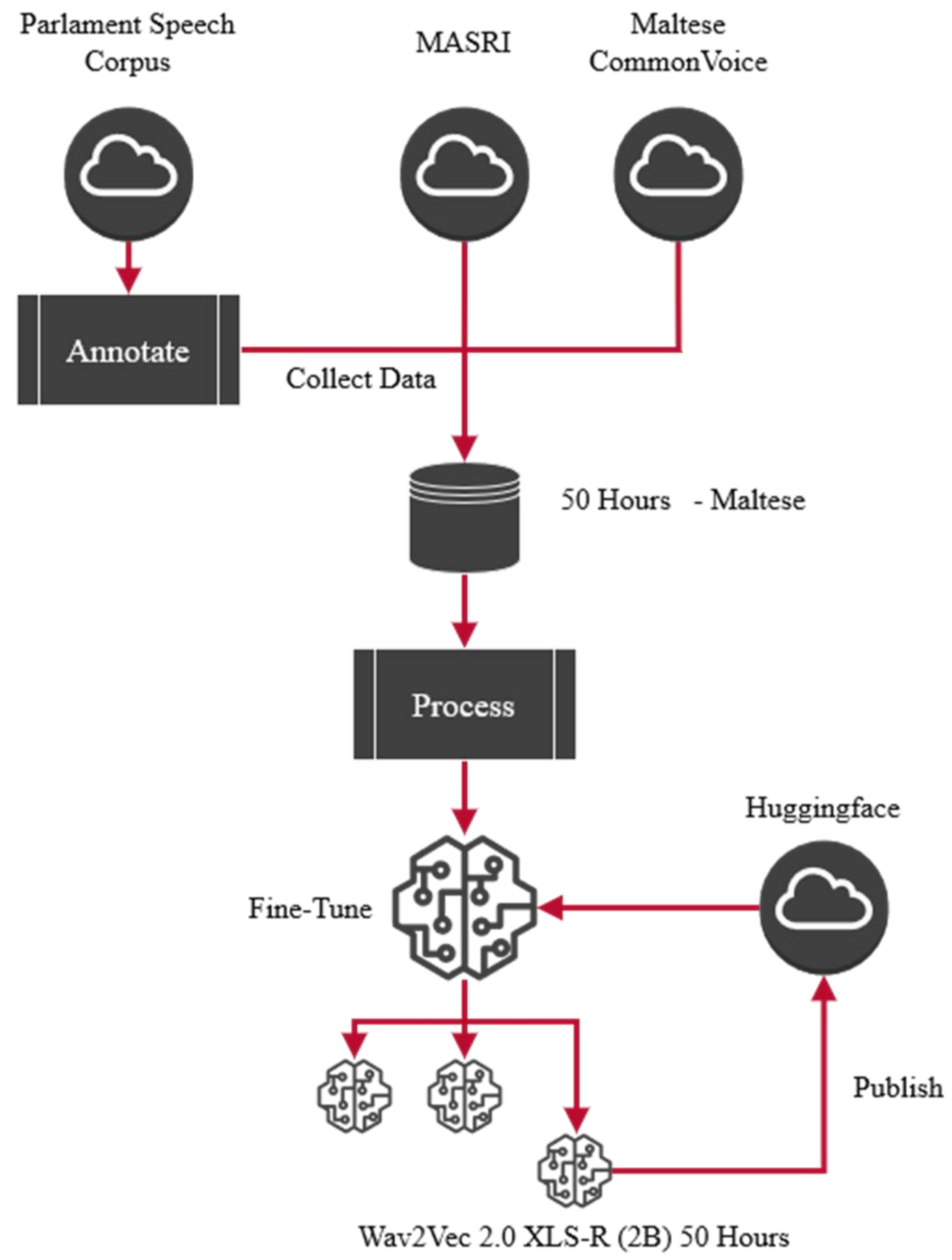
INTRODUCTION

Maltese is one of the low-resource languages for which the availability of labelled datasets for ASR training is scarce. This project presents a two-and-a-half-hour novel speech corpus, Parlament Speech. The audio is collected from Maltese parliamentary sessions and annotated at the word level. Traditionally supervised techniques, which are highly dependent on labelled training, have dominated the deep learning ASR field. Recently, self-supervision, a method which relies on unlabelled data for training, has been popularised in various AI fields, including speech systems. This approach has gained popularity in most of the state-of-the-art ASR systems developed to this day because of the more generous amount of unlabelled data available, even for low-resource languages.

AIM

The main aim of this project is to develop a Maltese ASR model using the Wav2Vec 2.0 and Whisper systems. Through the research carried out, the advances in ASR produced by the XLS-R model make it the ideal choice for low-resource ASR.

ARCHITECTURE DESIGN



METHODOLOGY

Collect a novel annotated corpus from the Maltese Parliamentary Sessions.

Combine the Parliament Speech corpus with other annotated Maltese corpora

Split the data and fine-tune XLS-R and Whisper models to investigate the effect data scaling has.

Evaluate the models on a set of Maltese speech scenarios.

RESULTS

A total of 30 Maltese ASR models are trained and are evaluated using WER and CER.

The performance of most models is improved when more data is used during training. On this set, the best performing model is the XLS-R 2B model when fine-tuned on 50-Hours of data. In general, the 2B models outperforms the other models. Models trained on 1-Hour or less did not produce any legible output. Performance improves with the scaling of more Maltese data, but diminishing returns are reported immediately.

The XLS-R 2B (50) has achieved state-of-the-art results of 8.53% WER and 1.93% CER when evaluated on a test set extracted from the CommonVoice dataset.

CONCLUSIONS AND FUTURE WORK

The main aim of this project was to develop a Maltese ASR model using the Wav2Vec system. Not only was this achieved, but the current state-of-the-art for Maltese ASR is presented along with the process and work behind it.

Future work would include the continuing annotation of the Parlament Speech corpus. The Wav2Vec 2.0 (2B) 50-Hour model may be used to further annotation efforts for the Maltese language.

REFERENCES

1. 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., and Auli, M. XLS-R: self-supervised cross-lingual speech representation learning at scale. CoRR, abs/2111.09296, 2021