Distilled Malaysian Whisper: Enhancing Low-Resource ASR with Knowledge Distillation

Husein Zolkepli[*] Aisyah Razak[†] Halim Shukor[‡]

Abstract

1 Introduction

2 Data

Data gathering process is essential for developing a robust speech recognition system tailored to the Malaysian context. The diverse linguistic landscape in Malaysia necessitates the collection of speech data across multiple languages, including Malay, Mandarin, Tamil, and English, with a particular focus on the integration of Singlish—a form of English widely spoken in Singapore. The data sources utilized in this study include YouTube and the IMDA Speech-to-Text (STT) corpus for Singlish.

2.1 Youtube

YouTube serves as a rich repository of publicly available speech data, providing a wide array of content in multiple languages spoken in Malaysia. The process of gathering data from YouTube involves

2.2 IMDA-STT

Singlish, a variant of English that incorporates elements from Malay, Mandarin, Tamil, and various Chinese dialects, is prevalent in Singapore but also understood in some Malaysian contexts. To incorporate Singlish into our speech recognition system, we utilized the IMDA Speech-to-Text (STT) dataset,

[*]husein@mesolitica.com
[†]aisyahrazak171@gmail.com
[‡]mhalimshukor@gmail.com

Technical Report.

3 Pseudolabeling Untranscribed Data

4 Postfiltering Pseudolabelled Data

5 Postprocessing

6 Knowledge Distillation

7 Results

8 Acknowledgement

We would like to express our gratitude to NVIDIA Inception for generously providing us with the opportunity to train our model on the Azure cloud. Their support has played a crucial role in the success of our research, enabling us to leverage advanced technologies and computational resources.

We extend our thanks to the wider research community for their valuable insights and collaborative discussions, which have greatly influenced our work. This paper reflects the collective efforts and contributions from both NVIDIA Inception and the broader research community.

9 Conclusion

References