Fine-tuning an Automatic Speech Recognition Model for a Canadian Indigenous Counselling Program

Emmanuel Olaniyanu[1], Zahra Moussavi[1,2]

[1] Biomedical Engineering, University of Manitoba, Winnipeg, Canada
[2]Eleectrical and Computer Engineering, University of Manitoba, Winnipeg, Canada

Abstract— **Automatic Speech Recognition (ASR) systems are programs designed to transcribe or identify spoken language. Most modern ASRs are created using End to End Neural Networks and are largely dependent on the quantity and quality of available speech training data. The lack of accented speech data can lead to poor ASR performance with niche accents and voice types. The ASR model presented in this paper is designed to work within an interactive VR counselling software for Canadian Indigenous youth, with an elder. This paper outlines the use of fine-tuning and other data processing techniques to minimize the Word Error Rate of our ASR model. These techniques provide valuable insight into data selection and processing.**

Keywords— **Speech Recognition, Data Science, Data Selection, Machine Learning**

I. Introduction

Automatic Speech Recognition (ASR) systems are programs designed to transcribe or identify spoken language. They have many applications and can be found used in voice recognition, speech transcription, language translation etc. In early years of development, traditional ASR systems were designed to parse speech in modules, transcribing speech sounds from phonemes to words and then to sentences. Today, most modern ASRs are created largely using machine learning and probability theory [1]. This makes an ASR's system's performance, like most machine learning algorithms, dependent on the data used to train the system.

The performance of ASR systems is evaluated based on a metric called the Word Error Rate (WER). The WER is the proportion of incorrectly transcribed words to total amount of transcribed words. It is usually represented as a percentage and a lower percentage indicates greater performance.

The goal of this project is to train an ASR system for use in a counselling and mental health support program. This project functions as part of a larger team project. The objective of the main project is to design a virtual reality counselling program primarily for Indigenous youth. The project has three parts namely, a virtual reality (VR) environment for the counselling to take place, an ASR system to transcribe the speech of the user and a Virtual Elder/Healer to interact with the user. This paper will focus on the training of the Indigenous counselling program's ASR system.

II. Methods and disscussion

When selecting data for any machine learning algorithm one of the most important aspects to consider is the representativeness of the data. This poses unique challenges in ASR model training due to the lack of speech data for niche accents and voice types. Studies have shown that ASR systems significantly increase in performance when trained with speech data that is representative of their target audience in gender, anticipated noise levels and accent [2] [3] [4]. Most publicly available ASR models combat this issue by using large amounts of variable speech data. This is called robust model design and results in moderate ASR performance across multiple voice types. The opposite of robust model design would be specific model design. Design for high specificity aims to maximize the ASR's performance for a particular voice type. The performance of this design type is dependent on how well the selected speech data represents the intended user's voice type.

For this project, both methods mentioned above, were combined in a process called fine-tuning. A pre-trained general English model was re-trained using additional user representative data. The ASR system that was used for this project was Mozilla Deep Speech. Audiobooks were converted to compatible speech corpora to fulfill the training data requirements for the ASR model. The audiobooks were selected based on their similarity to the voice type of the intended user. Representative vocabulary was acquired by selecting audiobooks written by Indigenous authors and containing Indigenous cultural terms. Anticipating low noise levels, a similar noise environment was simulated in the speech data. Indigenous accented speech data was acquired by selecting audiobooks that have Indigenous narrators. Gender differences in speech data has also been shown to result in higher WER [5]. Research also shows that the greatest gender difference in voices is in vocal pitch frequency. In order to minimize WER due to differences in vocal pitch, different models are trained with pitch specific speech data. The final program will then select a pitch specific model for the user based on the user's vocal pitch. The user will also have the option to select a general model which has been trained by all the available speech data. The functional interactive VR program and the performance of both ASR models will be shown as part of the final presentation. The contrast in model performance will provide incite to the importance of data quality versus quantity with regards to general speech data selection, and more specifically, vocal pitch. The results of this project shed light on techniques that can be used to fine-tune and adapt pre-trained ASR models to differing speech features. This reduces the cost of training ASR systems and minimizes transcription errors due to inequalities in training data.

Acknowledgment

This study is supported by MITACS in partnership with Myera Inc.

References

1. Hannun, A. *et al*, "Deep Speech: Scaling up end-to-end speech recognition", <i>arXiv e-prints</i>, 2014. doi:10.48550/arXiv.1412.5567.
2. Nagórski, Arkadiusz et al. "Optimal selection of speech data for automatic speech recognition systems." *Interspeech* (2002).
3. Chen, C., Hou, N., Hu, Y., Shirol, S., and Chng, E. S., "Noise-robust Speech Recognition with 10 Minutes Unparalleled In-domain Data", <i>arXiv e-prints</i>, 2022. doi:10.48550/arXiv.2203.15321.
4. Hinsvark, A., "Accented Speech Recognition: A Survey", <i>arXiv e-prints</i>, 2021. doi:10.48550/arXiv.2104.10747.
5. Feng, S., Kudina, O., Halpern, B. M., and Scharenborg, O., "Quantifying Bias in Automatic Speech Recognition", <i>arXiv e-prints</i>, 2021. doi:10.48550/arXiv.2103.151