Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number: 18

When mistaking “car tyre” for “Kaitaia” is not okay – Towards quantifying language bias in speech technology

Henry An

Louis Martin Chuo

Jesin James, Josh Bensemann

15/04/2022

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Henry An

ABSTRACT: There has been tremendous development in the field of Automated Speech Recognition (ASR) in recent years, however, many issues in ASR are yet to be addressed effectively. In particular, regarding accent and dialect recognition, improvements are necessary on interpretation accuracy, recognition error measurements, accent bias etc. Methods such as end to end modelling, multitask modelling or ensemble modelling provide potentials for the improvements. Comparing the performance of ASR systems between different dialects have found that New Zealand dialect generally performs worse than a general American dialect of English. Further investigation into the circumstances is needed. This report outlines findings on the relevant literature review in this area, which provides a potential research project. The research intent including the research questions to be answered in the research project and the project plan is also outlined.

1. Literature Review

The Field of Automated Speech Recognition (ASR) has made much progress in recent years by using new machine learning technologies. However, despite this or perhaps due to it this progress has not been uniformed for all types of speech with certain dialects and accents of speech being more accurately recognized. This section seeks to summarise the current state of speech recognition systems with regard to their accuracy between different dialects and accents, how technologies and techniques are employed to interpret accented speech and current research into new techniques for automatic recognition into accented speech.

1.1. The ASR model

An automatic speech recognition (ASR) system allows a computer to take the audio file or direct speech from the microphone as an input and convert it into the text, preferably in the script of the spoken language. An ideal ASR should be able to “perceive” the given input, “recognize” the spoken words and then subsequently use the recognized words as an input to another machine so that some “action” can be performed on [1]

Common machine learning techniques employed in ASR include artificial neural networks, support vector machines, Gaussian mixture models and hidden Markov models. Recent developments in deep learning have provided significant improvements in ASR performance. Deep learning will play an important role in the future of ASR [2], [3].

Accent and dialect recognition is an important part of ASR, which is similar to language identification and speaker identification. They all classify variable-length speech sequences to utterance-level posteriors to obtain accent, speaker, or language ID [4]. [5] proposed to evaluate a state-of-the-art automatic speech recognition (ASR) deep learning-based model, using unseen data from a corpus with a wide variety of labelled English accents from different countries around the world.

1.2. The Speech Recognition Systems

Automatic Speech Recognition (ASR) has provided many systems for speech recognition, aiming to allow natural communication between humans and computers via speech, where natural implies similarity to the ways humans interact with each other [6]. These speech recognition systems can help to create voice control for robots, which is both important and difficult [7].

[7] divided existing speech recognition systems into two main classes: (1) open-source and (2) close-source code. Examples for close-source software include: Dragon Mobile SDK, Google Speech Recognition API, Siri, Yandex SpeechKit and Microsoft Speech API. Examples for open-source software include: CMU Sphinx, Kaldi, Julius, HTK, iAtros, RWTH ASR and Simon.

Modern speech recognition systems typically utilise a mixture of Hidden Markov Models (HMM) and Deep Neural Networks (DNN). These replaced the traditional technology which used Gaussian Mixture models. Ongoing research into ASR systems include using new structures such the usage of Long Short Term Memory for the creation of an improved ASR system for the Sichuan dialect [8]

1.3. The challenges with existing ASR models

Over the past few decades, there has been tremendous development ASR from home automation to space exploration. Though commercial speech recognizers are available for certain well-defined applications like dictation and transcription, many issues in ASR like recognition in noisy environments, accent and dialect recognition, multilingual recognition, and multi-modal recognition are yet to be addressed effectively [3], [5], [9].

A major challenge in Automatic Speech Recognition (ASR) system is to handle speech from a diverse set of accents [9]. With the advent of tools such as virtual personal assistant such as Siri or the implementation of automatic captioning systems for videos, automated speech recognition technology has become widespread in recent years. It has become apparent that not all dialects and accents are equal in the accuracy that ASR technologies can interpret them [5]. Focusing specifically on English language ASR, it can be seen that the American variety of English, when measuring by word error rate, achieves the best accuracy results across a variety of different popular modern ASR systems [10]. Further investigation into different types of American English finds that the accuracy is not equal among groups within the US either with Caucasian American varieties achieving the highest accuracy rates [11].

The challenges with the existing ASR models regarding accent and dialect recognition include interpretation accuracy, recognition error measurements, accent bias etc. These are not unexpected. Much of the data available to sample is of the American variety, for example in the Vox Forge corpus over half the body is classified as American. As

the ASR software has likely been trained on a greater quantity of American English, it has superior performance when encountering American English again. Given the vast number of varieties of English, each with different number of available samples, it may be infeasible to simply sample English varieties such that a similar performance is obtained for all of them, in addition to considerations on performance overall.

Despite the efforts made by recent research work on improving the accuracy of ASR systems, such as [9], [12], [13], [14], [15], further improvements are necessary when handling accented/ dialectal speech of less resourced dialects in order to bring the performance of ASR systems on them into parity with more prominent dialects such as American. This importance will increase in coming years as technology relying on ASR systems becomes more common.

1.4. The possible improvements

There are possible improvements to the existing ASR models regarding accent and dialect recognition.

There are two main approaches to interpreting accented or dialectal speech, single model and dual model. A dual model system includes a model to detect which accent is being used in the speech. These systems require a large body of data to achieve good results, however this is often times not available for all dialects [16]. For the most accurate results it would be necessary to train the model on a highly specific dialect but the more specific the dialect the more restricted the body of samples. Thus, users' accents must be identified at a level which there is a sufficient quantity of data to train the model but also generic enough to reduce the effect on accuracy from dialectal variation.

A single model ASR system as it implies only trains one model but trains to be able to recognize which words are pronounced differently in different accents. Transfer learning and multitask learning were also found useful for spoken accent recognition tasks [13]. These typically utilise methods such as end to end modelling, multitask modelling or ensemble modelling.

Many recent research results may be potentially adopted for the future improvements. For example, specific predictions are made in [17] regarding approaches that might be taken to leverage sociophonetic knowledge to improve social dialect-recognition accuracy in ASR systems. [13] proposed an end-to-end accented speech recognition. A preliminary tool was proposed to capture speech deviations from standard English pronunciations as a way of supporting the learners to improve their reading proficiency [15].

1.5. Quantitative measures

The quantitative measurement for identifying errors is the main evaluation approach in current Speech Recognition Systems.

Current methods of measuring accuracy in speech recognition systems include measurements on a system level such as word error rate (WER), phrase error rate (PER) and character error rate (CER) [10], [11], [8]. These measurements are calculated from the recorded results of testing speech on different ASR systems. Measurements can also be taken and calculated on a sample level such as perplexity. The suitability of the measures depends on the context and purpose for taking the measurements. For example, the CER measurement is a more appropriate for measure for accuracy on Chinese Dialects than English since characters correspond to a whole syllable in Chinese while in English they are not directly linked to sounds. Comparing WER to PER, WER would likely require less preparatory work as the total number of words can be counted automatically while total phrases will need to be manually classified. However, PER may be considered more suitable for measuring the performance of an ASR system as in many situations, such as for search, the corruption of any words would require manual intervention. In addition, WER would bias the results to speech with greater word density.

1.6. Recognition of NZ English

Studies comparing the performance of ASR systems between different dialects have found that although the New Zealand dialect of English is not the worst performing dialect overall, it generally performs worse than a general American dialect of English. For example, when tested with the IBM Watson system NZ English had a word error rate for ~ 0.45 which was greatly lower than the 0.9 of Indian English but still significantly higher than the WER of 0.25 recorded for American English [10].

A potential shortfall of this result is the fact there is not one NZ accent. The general New Zealand dialect fall on a scale between Cultivated, which has traits more similar to British RP, and broad which has more distinctively local characteristics. In addition to this there are also distinct variations of English spoken in New Zealand such as the Southland dialect or the Māori English.

The weaker performance of ASR on New Zealand English when compared to American English is likely due to multiple reasons. Firstly, there is relatively small body of NZ English to work with. For example, in the VoxForge corpus only 2 percent of the total sample are classified as New Zealand English [10]. In addition to this New Zealand is a relatively small country. At the time of this writing New Zealand has a population of approx. 5 million which is 5 times

smaller than Australia, 13 times smaller than the UK and 66 times smaller than the US. Thus, it is less likely for NZ English to be prioritized for the purposes of training ASR models.

Despite these factors the performance of NZ English is similar to or even better than other non-American dialects. For example, in a study by [18] on the performance of the YouTube Captioning system on various dialects, the New Zealand dialect achieved a similar word error rate to New England English with an average WER of ~0.4 and a superior performance when compared to Scottish English which averaged a WER of ~0.55. As well as this the New Zealand dialect was found to outperform the closely related Australian dialect in a study by [5]. Considering the significantly larger population and economy of Australia, it is unlikely that this discrepancy was due to a greater sampling of New Zealand speech in the training data. Taking these results, there are likely factors specific to the dialects themselves that influence the accuracy of ASR systems. Further investigation to clarify the performance on New Zealand English is needed.

One limiting factor in the current research is that there is not necessarily a classification of the task being given when recording the sample. [17] shows that the type of task being given to the speaker has an impact on the performance of ASR systems. In the study it was found that while Yakama speech had the highest phrase error rate when sampled during a conversation, it had the lowest PER while sample when performing a reading task. Classifying the tasks being performed could give some insights into the performance of NZ Dialect of English.

2. Research Intent

Based on the literature review in the previous section, although there has been tremendous development in the field of Automated Speech Recognition (ASR), many issues in ASR are yet to be addressed effectively. In particularly, regarding the accent and dialect recognition, improvements are necessary on interpretation accuracy, recognition error measurements, accent bias etc. Methods such as end to end modelling, multitask modelling or ensemble modelling provide potentials for the improvement. Comparing the performance of ASR systems between different dialects have found that New Zealand dialect generally performs worse than a general American dialect of English. Further investigation into the circumstances of the performance of NZ dialect English is needed. We would like to answer the research questions outlined below in this research project.

2.1. Research Questions

Overall: can we improve an ASR model to be more accurate at recognising accented speech using linguistic analysis to provide bias compensation?

What is an appropriate quantitative measure for identifying errors in current Speech Recognition Systems?

- Research upon different current methods on evaluating speech recognition systems
- Determine the best suited method for quantifying the errors due to accented speech

How accurate are current accent-aware speech recognition systems in detecting accents?

- Research on any current open-source projects that we can use to build upon
- Find current results on different projects for our chosen accents to compare upon
- Find out the accents with the highest error rates
- Research on current methods for how ASRs are interpreting accents
- Research on other methods of quantifying accents that may have not been implemented on current ASRs
- See if these methods can be more accurate than the current ones

How can the accuracy of ASR systems be improved with regards to recognition of NZ English accented speech?

2.2. Project plan

The full details of the work plan will be decided at a later date. We intend to distribute the workload equally and to have a flexible work schedule in order to accommodate for differing obligations between the two project partners. Primarily we will work based around agreed upon deliverable deadlines and keep a high level of communication using tools such as discord.

3. Conclusions

With the recent improvements in technology such as the usage of Deep Neural Networks and Hidden Markov Matrices, automated speech recognition technology has made great improvements in accuracy. Despite this there is still room for improvement, in particular attention has been given to the bias of ASR systems to certain dialects/ accents. The New Zealand dialect of English is one dialect where there is potential for further improvement in speech recognition technology but also investigation into the causes of bias in speech technology.

Acknowledgements

The author would like to thank the supervisors Jesin James and Josh Bensemann for their support and guidelines in the process of this project.

References

[1] M. Malik, M.K., Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6), pp.9411-9457, 2021.
[2] E. Trentin, and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” *Neurocomputing*, 37(1-4), pp.91-126, 2001.
[3] J. Padmanabhan, and M.J. Premkumar, Machine learning in automatic speech recognition: A survey. IETE Technical Review, 32(4), pp.240-251, 2015.
[4] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, L., “The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods,” In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6918-6922, June, 2021.

[5] G. Cámbara, A. Peiró-Lilja, M. Farrús and J. Luque, “English Accent Accuracy Analysis in a State-of-the-Art Automatic Speech Recognition System,” arXiv preprint arXiv:2105.05041, 2021.
[6] V. Këpuska, and G. Bohouta, “Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx),” *Int. J. Eng. Res. Appl*, 7(03), pp.20-24, 2017.
[7] R. Matarneh, S. Maksymova, V. Lyashenko, and N. Belova, “Speech recognition systems: A comparative review”, 2017.
[8] W. Ying, L. Zhang, and H. Deng, “Sichuan dialect speech recognition with deep LSTM network,” *Frontiers of Computer Science*, 14(2), pp.378-387, 2020.
[9] A. Jain, V.P. Singh, and S.P. Rath, “A Multi-Accent Acoustic Model Using Mixture of Experts for Speech Recognition,” in Proc. Interspeech, pp.779-783, 2019.
[10] C. Ike, S. Polsley, and T. Hammond, “Inequity in Popular Speech Recognition Systems for Accented English Speech,” in proceedings of 27th International Conference on Intelligent User Interfaces, pp. 66-68, Mar 22, 2022.
[11] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition,” in Proceedings of the National Academy of Sciences, 117(14), pp.7684-7689, 2020.
[12] M.Y. Tachbelie, S.T. Abate, and T. Schultz, “Multilingual speech recognition for GlobalPhone languages,” *Speech Communication*, vol. 140, pp. 71–86, 2022.
[13] T. Viglino, P. Motlicek, and M. Cernak, “End-to end accented speech recognition,” in Proceedings of Interspeech, pp. 2140–2144, 2019.
[14] A. Jain, M. Upreti, and P. Jyothi, “Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning,” in Proceedings of Interspeech, pp. 2454-2458, 2018.
[15] B. Oyo, BM. Kalema, “A preliminary speech learning tool for improvement of African English accents,” in proceedings of International Conference on Education Technologies and Computers (ICETC), pp. 44-48, Sep 22, 2014.
[16] M. Najafian, and M. Russell, “Automatic accent identification as an analytical tool for accent robust automatic speech recognition,” *Speech Communication*, 122, pp.44-55, 2020.
[17] AB. Wassink, C. Gansen, and I. Bartholomew, “Uneven success: automatic speech recognition and ethnicity-related dialects,” *Speech Communication*. Mar 12, 2022.
[18] R. Tatman, “Gender and dialect bias in YouTube’s automatic captions,” in Proceedings of the first ACL workshop on ethics in natural language processing, pp. 53-59, April, 2017.