Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number:18

Addressing Accent Bias in Speech Recognition Systems

Louis Chuo

Henry An

Jesin James, Josh Bensemann

12/04/2022

Declaration of Originality

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

Louis Chuo

Name: Louis Chuo

ABSTRACT: This report summarizes the current literature in the field of considering accent bias in speech recognition systems. Current systems and their performance with accented speech is reviewed and considered, as well as the current literature surrounding potential solutions. Finally, we develop our Research Intent by considering the literature, and break this down to Research Questions and Objectives to further break down the steps needed for answering our question.

1. Literature Review

Current modern societies have increasingly relied on Speech Recognition technology for tasks such as speech-to-texts and with virtual assistants such as Apple's Siri. Artificial intelligence is used for these to take voices and process them into understandable information by the processing machine to make use of. Due to human nature, there will be variations on the pronunciations and speech patterns of different people, which can lead to speech recognition systems not being as accurate as the dataset that they are being trained do not account for these accents. As such, research has and is being done to try and accommodate to this and allow for a more accurate speech recognition system.

The focus for this literature review is to answer these questions pertaining to the topic of speech recognition accent bias: "How is accent bias quantified in modern speech recognition systems?", "How are current Speech Recognition systems performing against these biases?", and "What are the current trends and solutions for solving the issue of accent bias in speech recognition software?"

1.1. Current systems & their performance

Research has been done to determine if accent affects the performance of speech recognition systems. Huang et al. [1] found out that of all the variabilities in speech, accents and gender are the most important to consider. As such, further research was done to determine what the exact effects are as well as the initial steps to try and solve these biases. Research by Huang et al. [2] showed that current speech systems showed a 40-50% error increase when trying to classify an accent with a model trained by a different one. They have determined that there are two methods that can be done as a solution to these problems. First, pronunciation adaptation can be used, where the ASR system adapts specifically on the pronunciation of the speech, tracking a user's speaking speed and style, and modifying the system to account for these, usually by modifying the neural network layers to account for the differences between the accent and the standard English accent. The second is a more general model that is dependent on the current accent, with the usage of automatic accent identification, classifying the accents based on its features and using the appropriately trained recognition model to

translate. This can be a problem on more generalized ASRs, as the number of accents that can be translated accurately is determined by the amount of pre-trained models for those specific accents.

More research was done to determine if automatic accent identification (AID) helps in the accuracy of detecting accents. Najafian & Martin [3] used an i-vector based AID to determine its accuracy on British-accented speech. I-vectors are a compact way of determining and quantifying the unique utterances of speech. Each speech pattern has a different i-vector value, and accents have very similar values and can be used to group together similar ones. This can be utilized to sort out speech in the model and classify them based on their accents, grouping the ones that are the most like each other. They used the ABI-1 language corpus, which consists of British-English accents such as Irish and Scottish, which are all labeled for training the model. They used models trained with the given accents, then used them depending on how the speech input is classified. This proved successful, as the overall average Word Error Rate for the corpus decreased, showing that the classification of accents is successful. They also concluded that the information obtained from the i-vectors in the experiments can be used to augment current speech recognition neural networks to better perform against accents dynamically, instead of using pretrained models.

Ying et al. [4] researched on the phonetic adaptation of ASRs, using the Sichuan dialect of Mandarin, collecting their own dataset, and building a lexicon of common phoneme sequences in the Sichuan dialect. They combined a hidden Markov chain (HMM) together with a long short-term memory network (LSTM) as this can overcome the limitation of a normal deep neural network that can only capture the context of a fixed amount of information, which can hinder the recognition of phonetic feature of an accent. This proved to be successful, producing a character error rate of 18.09% which is the lowest they know of currently on the specific accent. As the Mandarin language uses characters instead of words, character error rate was used in lieu of word error rate, but they are both functionally the same. Furthermore, Wassink et al. [5] conducted research on more phonetic variations on languages, this time specifically on a multi-ethnic sample of pacific-northwest American English accents. They utilize socio-phonetic methods to further identify and classify the different vocalizations of their dataset, to allow for a more accurate quantitative representation of how each accent diverges from the standard American English accent. They then used Microsoft's speech Recognition software to test out the error rates, comparing the translated speech to its correct text by hand, to accurately determine the exact errors that are produced while translating. They determined that around 20% of the errors shown are caused by these specific socio-phonetic features such as mispronunciations and devoicings. Moreover, a majority of the errors detected were from

speech inputs that use conversational language, compared to lexical tasks and reading passages, implying that pronunciation is not the root cause of word errors. These systems can be leveraged, specifically the consideration of fine phonetic details, in further improving speech recognition accuracy, although this research implies that word error rates may not be the most accurate representation of pure phenetic accent errors on ASR systems.

Current commercial Speech Recognition systems do not yet account for accent bias. Koeneke et al. [6] examined five state-of-the-art ASR systems: Amazon, Google, IBM, and Microsoft, using the Corpus of Regional African American Language (CORAAL) to see if there is specifically racial bias on these systems. Racial bias can be equated to accent bias as both are similarly have socio-phonetic features that are divergent from the standard data that commercial speech recognition systems have. They have found that the average word error rate for black American speech is considerably worse than for white American speech, consistently through all the five ASRs that they used, showing racial bias leaning towards white American speech. They also do point out that, as their testing methods use the Word Error Rate, text-to-speech was used to compare the recognized data to the source. Although ASR systems are trained to translate with a wide vocabulary, there is a chance that they simply mistranslated slang or uniquely structured grammar, thus WER may not be the most accurate representation of accented speech recognition.

Tatman [7] also did similar research, using YouTube's automatic captions and evaluating their accuracy in terms of the video's tagged country as a reference for their accent. YouTube's automatic captions use Google's ASR system, and YouTube itself has a large userbase making it ideal for testing. Accents from California, Georgia, New England, New Zealand, and Scotland were sampled, and the videos that were sampled were on the topic of the "accent challenge" where people were to recite a list of words as a "challenge" to see the effects of various accents. This meant that ASRs could not make use of language models that try to contextualize sentences, making sure that the system will only recognize the words itself. The results in accent WER showed that Californian had lower overall WER than the other accents, further solidifying that current commercial ASRs are biased towards white American accent. More importantly, they also found that the New Zealand English accent have a confidence interval that is unusually high in comparison to the others, suggesting that New Zealand English itself has a high degree of variability.

1.2. Current trends & solutions

Current research is being done into potentially combining accent recognition with the improvement of the recognition itself by identifying and accounting for the deviations from the standardized English accent. Multiple research papers

have been published [8, 9, 11] that follow this logic, using, two separate neural network models to allow for flexible speech bias learning while also being able to do accent recognition. Oyo & Kalema [8] used African English as the base dataset to develop a dual speech recognition engine. This works by firstly using a speech classifier engine to segregate the different accents, then a second recognizer is used to classify the differences between the current accent and the standard English accent. This difference is then used in a tool that can be used to further improve a person's pronunciation by providing feedback to the person on the utterances that have the most deviation to the standard accent. Quantifying the difference between current accents and the standard English accent, as well as classifying this in the basis of its accent can be extremely helpful in allowing a multi-accent ASR system to dynamically recognize accents without having to be explicitly trained by the current accents being detected.

Jain et al. [9] expands further on this research. A model like the dual-model framework is used, this time with the first model being an accent classifier as well as phone recognition to translate speech to text. This is trained in conjunction with another model that classify the differences from the detected accent to the standard English accent. This then produces accent embeddings that can be used as auxiliary inputs to the phone recognition model to allow for accented speech to be transcribed more accurately. Moreover, they also used a different, more robust dataset, the Mozilla Common Voice Corpus [10] which compiles over 14,000 validated hours of different voices. It also has a wide variety of accents, which means that accent recognition can be more robust. The research uses the most well represented accents in the datasets, as well as using the model to try and detect accents that are not used to train the model, labeled as "unseen" accents. Using the accent embeddings on the model achieved a 15% overall WER reduction on the trained accents and a 10% WER reduction on the untrained accents, showing promise in the field to further develop and improve the system. Unseen accents, even if it was improved on the accuracy overall, was still very inaccurate, with one of the unseen accents, South Indian English, still having an average WER of around 50%, thus, there is still room for improvement with the model.

Viglino et al. [11] improves upon the prior's work, using the end-to-end ASR system to improve the model and produce more accurate results. The end-to-end ASR system is based on the Deepsearch 2 architecture, developed by Amodei et al. [12] which was developed to recognized between English and Mandarin speech, as such, it will be more accurate in classifying accents than other systems. They also found, through testing, that NZ English is consistently more accurate compared to South Indian English as unseen accents, implying that the test accents are closer in features to the NZ accent.

In particular, the accent features of the NZ accent are within the parameters of the British, Australian, Canadian, and American English accents, therefore, these languages can be tested alongside NZ English on further work to allow for more comparisons.

1.3. Summary & Potential Research Avenues

It was shown that the New Zealand accent has a high degree of variability in comparison to other English accents, and there is very little research done to classify and solve this issue, this may be a potential avenue for potential research. There is also potential research on combining the current multi-framework model being researched with the socio-phonetic analysis to improve upon the accuracy of accent translation further, allowing for a clearer definition of the differences on untrained languages to better detect these. Lastly, word error rate, which is the most common measure for ASR accuracy has been shown to potentially not be the most accurate specifically on quantifying errors form accent purely, as such, further research can be made to determine if there is a better measure.

2. Research Intent

Based off our literature review, we have decided on our project's overall question: Can an ASR model be improved to be more accurate at recognizing New Zealand English accented speech using linguistic analysis to provide bias compensation?

Furthermore, we broke this down to more specific questions to help in identifying the objectives in answering this question:

- What is an appropriate quantitative measure for identifying errors in current Speech Recognition Systems?
- How accurate are current accent-aware speech recognition systems in detecting accents?
- How can the accuracy of ASR systems be improved with regards to recognition of NZ English accented speech?

Based off this, we developed Research Objectives to help us in answering these questions:

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

These tasks will be split with my partner and I, and we prefer not to concretely assign the tasks, in the case of blockages in progress with each other, as well as in dependence to how busy we are, as we have varying schedules to each other.

3. Conclusion

This paper is a review of the status of Automatic Speech Recognition systems pertaining to its performance with accents. Current methods to identify and quantify this are discussed. The performance of the currently commercially available Speech Recognition systems is also considered for reference to the improvement of its accuracy. This literature review also considers the current trends that attempt to solve the accent affect on the accuracy of ASRs, with multitask Neural Network frameworks, and outlines potential avenues of further research derived from the current literature. Finally, we state our Research intent by formulating research questions that break down the steps needed and derive research objectives from these to specify the tasks needed to attempt in answering our research questions.

References

[1] C. Huang, T. Chen, S. Li, E. Chang, and J.L. Zhou. "Analysis of speaker variability." In *INTERSPEECH*, pp. 1377-1380, 2001.

[2] C. Huang, T. Chen, and E. Chang, “Accent issues in large vocabulary continuous speech recognition.” *International Journal of Speech Technology*. 7(2):141-53, April 2004.

[3] M. Najafian, and M. Russell. "Automatic accent identification as an analytical tool for accent robust automatic speech recognition." *Speech Communication* 122: 44-55, 2020.

[4] W. Ying, L. Zhang, and H. Deng. "Sichuan dialect speech recognition with deep LSTM network." Frontiers of Computer Science 14, no. 2: 378-387, 2020.

[5] A.B. Wassink, C. Gansen, and I. Bartholomew. "Uneven success: automatic speech recognition and ethnicity-related dialects." *Speech Communication*, 2022.

[6] 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." *Proceedings of the National Academy of Sciences 117*, no. 14: 7684-7689, 2020.

[7] 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. 2017.

[8] B. Oyo, and B.M. Kalema. "A preliminary speech learning tool for improvement of African English accents." In *2014 International Conference on Education Technologies and Computers (ICETC)*, pp. 44-48. IEEE, 2014.

[9] A. Jain, M. Upreti, and P. Jyothi. "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning." In *INTERSPEECH*, pp. 2454-2458. 2018.

[10] Mozilla. Common voice. [Online]. Available: https://commonvoice.mozilla.org/

[11] T. Viglino, P. Motlicek, and M. Cernak. "End-to-End Accented Speech Recognition." In *INTERSPEECH*, pp. 2140-2144. 2019.

[12] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in ICML’16: Proceedings of the 33rd InternationalConference on International Conference on Machine Learning - Volume 48, M. F. Balcan and K. Q. Weinberger, Eds. JMLR.org, 2015