Abstract No: SP-01

Comparative study of neural network based speech recognition algorithms for Sri Lankan accent

J. K. D. R. Jayasekara[1], M. V. M. Jayathilake[2*] and S. V. S. Gunasekara[1]

[1]Department of Information Technology, CINEC Campus, Sri Lanka
[2]Sri Lanka Institute of Advanced Technological Education
mekala@sliate.ac.lk*

Speech recognition is a technology which involves processing and interpreting human speech into a written format. Advancements in technology have led to the development of sophisticated speech recognition algorithms with the use of neural networks, machine learning and artificial intelligence. Automatic Speech Recognition (ASR) is being used worldwide for developing various applications including automated devices to communicate with humans such as Alexa, Siri and Artificial Intelligence Chatbots. Several machine learning algorithms, Natural Language Processing (NLP) techniques, Hidden Markov Models (HMM) and neural networks are used to create the foremost speech recognition systems. However, most speech recognition algorithms are yet to overcome the many barriers which come along with the technology. Variations in pronunciations and accents, lack of fluency, speech clarity, speed of speech and language technicalities are just few of the challenges faced by modern day speech recognition algorithms. These problems are magnified in lesser-known languages and accents. The purpose of this research is to compare the accuracy of multiple speech recognition systems for unexplored accents such as the Sri Lankan accent. A comparison was conducted between three leading neural-network based speech recognition systems regarding their accuracy in recognizing speech spoken in a Sri Lankan accent. The primary objective of this study was to determine the system which applies the most efficient algorithm for recognizing speech with language nuances. Google Cloud Speech-to-Text, Mozilla DeepSpeech and CMU Sphinx were the three systems used in the research comparison. Quantitative secondary data was used to analyse existing speech recognition systems and their accuracy in interpreting speech in English accents. Furthermore, experimental research was conducted using primary audio data gathered using different speakers. Six selected sentences were converted to a verbal format in the form of individual audio files in the .wav format. Two versions of Sri Lankan accents were recorded for each sentence. An algorithm was designed in the Python language to calculate the Word Error Rate (WER) for each system and determine the one with the lowest error rate. Word Error Rate is a metric used to calculate the accuracy of text transcribed by speech recognition systems. The mean WERs obtained were 0.86, 1.05 and 0.59 for Mozilla DeepSpeech, CMU Sphinx and Google Cloud Speech-to-Text respectively. While the results provide conclusive evidence that the Google Cloud ASR system is the best at identifying speech in a Sri Lankan accent, it could be clearly observed that all three systems encountered difficulties when recognizing homophones and words with contradictory pronunciations. The outcome of this research indicates that although speech recognition systems have had major improvements over the years, there are still a lot more enhancements to be done in order to provide accurate and efficient speech-to-text transcriptions. The systems should be trained with larger and miscellaneous datasets which include speech from diverse languages and accents. As of now, the Google Cloud speech recognition system displays optimal performance when interpreting speech in Sri Lankan accents.

Keywords: Google Cloud Speech-to-text, Neural networks, Speech recognition, Sri Lankan accent, Word Error Rate