A Survey of Speech Recognition for People with Cerebral palsy

Yu-Ru Wu, Jason C. Hung, and Jia-Wei Chang

National Taichung University of Science and Technology, Taichung City, Taiwan
wu870626@gmail.com

Abstract. This study aims to address the communication barriers related to speech for individuals with cerebral palsy, with the goal of using technological methods to assist or alleviate difficulties in oral communication. To achieve this, the study plans to analyze and test mainstream speech recognition services or platforms available in the market to understand their current speech recognition capabilities for individuals with cerebral palsy, and explore the possibility of assisting them in solving their communication problems, in order to enhance their quality of life and promote their social skills. As the author is a person with congenital cerebral palsy, the study is particularly meaningful to him because the congenital brain damage affecting the nervous system has made his speech unclear, seriously affecting his ability to express himself orally. Therefore, the author plans to record a dataset of speech samples from individuals with cerebral palsy, collecting conversations from various aspects of daily life. This dataset will be tested and analyzed using mainstream speech recognition services such as Google, Microsoft, and YaTing, among others, in order to infer the current difficulties in speech recognition technology for individuals with cerebral palsy and propose potential solutions for oral communication barriers, with the hope that the contribution of this research will promote the development of mature assistive technologies for individuals with communication difficulties in the near future.

Keywords: Congenital Cerebral Palsy, Speech Recognition, Speech Clarity.

1 Introduction

This study intends to record speech files of cerebral palsy patients and collect the correct textual answers of their speech, in order to provide a self-made speech dataset for cerebral palsy patients. Meanwhile, using the speech recognition services provided by well-known artificial intelligence companies such as Google, Microsoft, and YaTing as the testing benchmark, the study aims to investigate and analyze whether the most advanced speech recognition technology can recognize the speech of cerebral palsy patients, with the goal of assisting or alleviating their speech communication barriers.

2 Related Works

In the early 1960s, speech recognition technology was primarily based on pattern matching methods. Pattern matching involved comparing input speech to pre-stored speech templates to determine speech content. Due to the complexity of speech variations, the accuracy of this method was limited. However, contemporary speech recognition systems can only recognize basic single-speech sentences. In the late 1970s, a method based on Hidden Markov Models (HMM) emerged. This method built state transition models and mixtures to achieve higher accuracy than pattern matching, but still had certain limitations. The quality of speech recognition may be affected by environmental factors, such as noise, due to the characteristics of the speech signal itself [1].

In the 1990s, neural network methods emerged as a type of speech model that is capable of learning and adapting. This method can be trained using backpropagation algorithm, and its accuracy is higher than the previous two methods. However, it requires more computing resources and data. For example, [2] proposed a time-reversal backpropagation neural network speech recognition method.

In the 2000s, with the development of deep learning, speech recognition methods based on deep learning emerged. Deep learning builds deep neural networks to learn speech features, which further improves speech recognition accuracy. Convolutional neural networks and recurrent neural networks are commonly used models, for example, [3] proposed a speech recognition method based on recurrent neural networks that converts speech signals into text sequences.

In recent years, with the continuous development of speech recognition technology, new methods based on deep learning have emerged, such as end-to-end learning, which directly maps speech signals to text sequences and avoids the complexity of intermediate steps, further improving the speech recognition performance [4]. In general, the technology of speech recognition has been advancing constantly, evolving from pattern matching, hidden Markov models, neural networks, to deep learning and end-to-end learning methods. These advancements have continuously improved the accuracy and application range of speech recognition.

In the field of speech, the speech model plays an important role. The knowledge base behind the speech model makes predictions based on the context of the knowledge base, providing appropriate sentences. By pre-training the model on a large amount of text, performance on many downstream tasks often improves with different model sizes and increasing amounts of unsupervised data using transfer learning. The model does not need to learn external knowledge, only to memorize, and can provide appropriate responses in a speech question-and-answer format [5].

In summary, early speech recognition technology was developed based on pattern matching, until the current speech recognition technology was developed using deep learning-related techniques.

3 Methodology

3.1 Metrics of Speech Recognition

For the analysis and ranking of speech recognition results for people with cerebral palsy, three speech models, Google, Microsoft, and YaTing, were used for recognition. Using the Character/Word Error Rate (CER/WER) scoring tool, the accuracy of the speech recognition models was analyzed based on the recognition results, and the other commonly used evaluation indicators for speech recognition, such as Match Error Rate (MER), Word Information Preserved (WIP), and Word Information Lost (WIL), were used to obtain the average, median, maximum, minimum, and standard deviation for each evaluation [6].

1. CER represents the indicators of word error rate. It calculates the total number of hits, insertions, deletions, and substitutions of words in the recognized sentence compared to the correct reference sentence, and divides it by the total number of words in the reference sentence to obtain the word error rate percentage.
2. MER is a measurement metric in speech recognition that represents the percentage of characters in the recognized speech that are missing compared to the correct speech. A lower MER value indicates better integrity in speech recognition, and the value range of MER is between 0 and 1.
3. WIP is an evaluation indicator in speech recognition that measures the percentage of correct word count in the recognized text by the speech model, compared to the total word count in the correct transcription. A higher WIP indicates better recognition performance of the speech recognition system. The WIP value ranges from 0 to 1.
4. WIL measures the text loss rate of a speech recognition model. It is the percentage of incorrectly recognized words in the total number of correct words. A lower WIL indicates better recognition performance of the speech recognition system. The WIL value ranges from 0 to 1.

3.2 Speech Recognition Performance of Google, Microsoft and YaTing

Speech recognition technology has come a long way in recent years, and major players like Google, Microsoft, and YaTing are leading the charge in delivering accurate and efficient speech recognition services. Therefore, we conduct the experiments with the three service providers. Their performance of CER, MER, WIL and WIP are shown in the Figure 1.

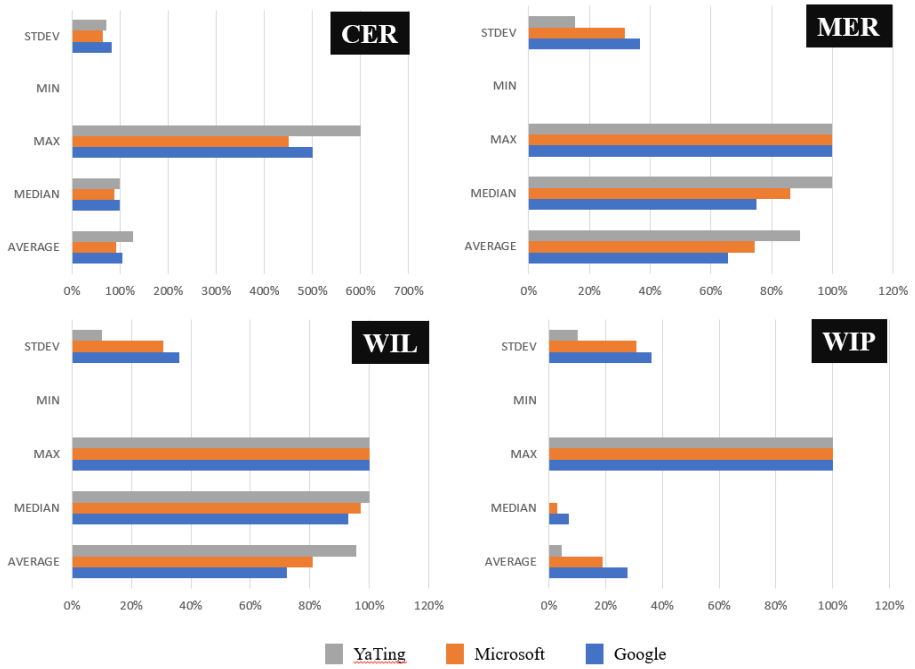

Fig. 1. Comparisons of YaTing, Microsoft and Google on CER, MER, WIL and WIP.

1. The experimental results of Google speech recognition are as follows. The average CER is 104%, with a median of 93%, a maximum of 500%, a minimum of 0%, and a standard deviation of 83%. The average MER is 66%, with a median of 75%, a maximum of 100%, a minimum of 0%, and a standard deviation of 37%. The average WIP is 28%, with a median of 7%, a maximum of 100%, a minimum of 0%, and a standard deviation of 36%. The average WIL is 72%, with a median of 93%, a maximum of 100%, a minimum of 0%, and a standard deviation of 36%.
2. The experimental results of Microsoft speech recognition are as follows. The average CER is 93%, with a median of 89%, a maximum of 450%, a minimum of 0%, and a standard deviation of 65%. The average MER is 74%, with a median of 86%, a maximum of 100%, a minimum of 0%, and a standard deviation of 32%. The average WIP is 19%, with a median of 3%, a maximum of 100%, a minimum of 0%, and a standard deviation of 31%. The average WIL is 81%, with a median of 97%, a maximum of 100%, a minimum of 0%, and a standard deviation of 31%.
3. The experimental results of YaTing speech recognition are as follows. The average CER is 126%, with a median of 100%, a maximum of 600%, a minimum of 0%, and a standard deviation of 73%. The average MER is 89%, with a median of 100%, a maximum of 100%, a minimum of 0%, and a standard deviation of 15%. The average WIP is 5%, with a median of 0%, a maximum of 100%, a minimum of 0%, and a standard deviation of 10%. The

average WIL is 95%, with a median of 100%, a maximum of 100%, a minimum of 0%, and a standard deviation of 10%.

The CER represents the accuracy rate, and a lower rate indicates better recognition capability. Among the four evaluation indicators, the average CER of Microsoft's speech model is 93%, compared to Google's 104% and YaTing's 126%. The median CER of Google is 93%, Microsoft is 89%, and YaTing is 100%. The maximum CER of Google is 500%, Microsoft is 450%, and YaTing is 600%. The standard deviation of Google's CER is 83%, Microsoft's is 32%, and YaTing's is 73%. Therefore, Microsoft's speech model performs better in terms of CER recognition ability. Ranking of average CER is Microsoft > Google > YaTing.

The MER represents the word omission rate, and a lower rate indicates better speech recognition capability. Among the four evaluation indicators, the average MER of Google's speech model is 66%, while Microsoft's is 74%, and YaTing's is 89%. In comparison, the average MER of Google is much lower than that of Microsoft and YaTing. The median MER of Google is 75%, Microsoft is 86%, and YaTing is 100%. The median order is Google, Microsoft, and YaTing. The maximum and minimum values of the three speech models are all 100% and 0%, respectively. The standard deviation of Google's MER is 37%, Microsoft's is 32%, and YaTing's is 15%. The standard deviation order is YaTing, Microsoft, and Google. The order of average MER is Google > Microsoft > YaTing.

WIP represents the percentage of unidentified speech in the total amount of speech, and a higher value indicates better speech recognition ability. Among the four evaluation indicators, in terms of WIP, the average value for Google is 28%, for Microsoft is 19%, and for YaTing is 5%. The order of average values is YaTing, Microsoft, and Google, respectively. The median values for WIP are 7% for Google, 3% for Microsoft, and 0% for YaTing, and the order of median values is Google, Microsoft, and YaTing, respectively. The maximum and minimum values for the three speech models are 100% and 0%, respectively. The standard deviation for Google WIP is 36%, for Microsoft WIP is 31%, and for YaTing WIP is 10%. The order of standard deviation values is Google, Microsoft, and YaTing, respectively. The average WIP values are in the order of Google > Microsoft > YaTing.

WIL represents the percentage of identified speech in the total amount of speech, and a lower value indicates better speech recognition ability. Among the four evaluation indicators, in terms of WIL, the average value for Google is 72%, for Microsoft is 97%, and for YaTing is 95%.The order of average values is Google, Microsoft, and YaTing, respectively. The median values for WIL are 93% for Google, 97% for Microsoft, and 100% for YaTing, and the order of median values is Google, Microsoft, and YaTing, respectively. The maximum and minimum values for the three speech models are all 100% and 0%, respectively. The standard deviation for Google WIL is 36%, for Microsoft WIL is 31%, and for YaTing WIL is 10%. The order of standard deviation values is Google, Microsoft, and YaTing, respectively. The average WIL values are in the order of Google > Microsoft > YaTing.

Based on the above four evaluation indicators and the testing data of 500 speech files, 500 sentences were selected for the evaluation and the speech recognition ability of the 500 sentences was sorted. The order of speech recognition ability is Google >

Microsoft > YaTing. Each of the three speech recognition services, Google, Microsoft, and YaTing, has its own advantages. Google performs better in MER, WIP, and WIL, while Microsoft performs better in CER. However, the speech recognition performance of the models may vary depending on the speech data used. The experiments only represent the results of this small-scale evaluation for people with cerebral palsy.

References

1. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
2. Lippmann, R. P. (1997). Speech recognition by machines and humans. Speech communication, 22(1), 1-15.
3. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649).
4. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.
5. Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model?. arXiv preprint arXiv:2002.08910.
6. Xu, B., Tao, C., Feng, Z., Raqui, Y., & Ranwez, S. (2021). A benchmarking on cloud based speech-to-text services for french speech and background noise effect. arXiv preprint arXiv:2105.03409.