The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Accelerating Hakka Speech Recognition Research and Development Using the Whisper Model

Ching-Yuan Chen
Department of Cultural Creativity and Digital Marketing, National United University
minicr1234@gmail.com

Yun-Hsiang Hsu
Department of Cultural Creativity and Digital Marketing, National United University
yuripeyamashita@gmail.com

Chen-Chi Chang*
Department of Cultural Creativity and Digital Marketing, National United University
kiwi@gm.nuu.edu.tw
*Corresponding author

Abstract

Language preservation has become increasingly urgent with globalization and rapid technological advancement. Minority languages, such as Hakka, are particularly vulnerable. This study aims to expedite the research and development of Hakka speech recognition by integrating Open AI's Whisper model with online Hakka speech resources. This study developed an efficient speech recognition model for Hakka, a low-resource language, and provided insights into its applications for preserving and popularizing Hakka culture. This paper addresses the challenge of building an Automatic Speech Recognition (ASR) model for the Hakka language. Utilizing Open AI's Whisper technology, this study presents a complete workflow for training and deploying a Hakka language ASR model. The end product could be a vital tool for digital Hakka language education and intelligent living applications.

Keywords: Automatic Speech Recognition, Hakka Speech Recognition, Whisper Model, Minority languages

1 Introduction

The preservation and promotion of minority languages like Hakka are becoming more crucial. One way to promote a language's usage and preservation is through technology. This paper explores the potential of using Open AI's Whisper model to accelerate the development of a Hakka speech recognition system. By combining the Whisper model with online Hakka speech resources, this study seeks to create an effective, scalable speech recognition model for this low-resource language. For numerous languages globally, there is an insufficient amount of annotated speech data to train an Automatic Speech Recognition (ASR) model effectively (Scharenborg et al., 2017). One of the most considerable challenges in Hakka language AI technology is developing an efficient Hakka Language ASR model. Achieving this would advance digital education in the Hakka language substantially. Not only could such a system evaluate the pronunciation accuracy of Hakka words and sentences, but it could also serve as an aid in teaching Hakka pronunciation. This paper discusses constructing a Hakka language ASR model using Whisper, an ASR neural network model provided by OpenAI (Radford et al., 2023), and online resources such as the Hakka Language Database, YouTube videos, and online education platforms.

2 Database Profile

The database employed in this research constitutes a diverse collection of sources, encompassing the Hakka Language Database, YouTube videos, online educational materials, multimedia news content, and electronic dictionaries. This eclectic assortment of resources was chosen to construct a robust and all-encompassing dataset, affording the model the opportunity to glean insights from various accents, pronunciations, and lexical contexts. For the training phase, this study made use of data contributed by the 2023 Formosa Speech Recognition Challenge (Hakka ASR),

specifically the FSR-2023-Hakka-Lavalier-Train dataset, in conjunction with a Hakka speech dataset licensed by the Hakka Affairs Council. These sources provided a cumulative total of 80 hours of Hakka speech data. Furthermore, to enhance the training model, the study proactively collected additional Hakka speech data from various online platforms, including educational websites, YouTube, Podcasts, and news articles.

3 Training Procedure

Our study employed the Whisper large-v2 model for training. In the process of gathering speech data, there are alternative techniques that derive some of their foundation from the research presented in the O-COCOSDA 2020 paper authored by Dr. Hung-Shin Lee and his colleagues (Chen et al., 2020). This paper serves as a significant reference point in understanding the nuances of these methods. In the training procedure, the first step involves data preparation, where audio files and their corresponding sentences are collected to form a diverse dataset that includes a wide range of voice and language patterns. This is followed by formatting the dataset, where the gathered audio files and text labels are organized into a structured format like CSV or PyTorch's DataLoader. To ensure consistency and facilitate the model's learning process, the sample rate of all audio files is converted to 16KHz. Utilizing WhisperFeatureExtractor, padding is applied to audio files shorter than 30 seconds to make them uniform in length, and log-Mel spectrograms are generated to capture the characteristics of the audio signals. Subsequently, WhisperTokenizer is used to convert the model's output tokens into human-readable text (for example, [1169, 3797, 3332] -> “the cat sat”). The feature extractor and the tokenizer are then bundled together to form a WhisperProcessor, which streamlines the model's usability in subsequent steps. Two metrics are used for validation: Word Error Rate (WER) for word-level error and Character Error Rate (CER) for character-level error (Shah et al., 2022). Specific training parameters are set, which include the learning rate, batch size, and gradient accumulation steps. The model is then trained using the prepared settings and dataset. Upon successful training, the model becomes ready for deployment in various applications, including voice assistants and automated transcription services (Gandhi, 2022). The specific steps included:

(1) Data Preparation: Aligning sentences with corresponding audio files

(2) Inputting the dataset and converting it into a dataset format

(3) Resampling the audio files to a 16KHz rate

(4) Implementing the WhisperFeatureExtractor

- Padding audio files shorter than 30 seconds to reach 30 seconds

- Converting the audio into Mel-spectrograms

(5) Implementing the WhisperTokenizer

(6) Assembling the WhisperProcessor for simplified usage

(7) Setting up validation metrics: WER for Hakka characters and CER for Hakka phonetics

(8) Configuring training parameters

(9) Training commencement

(10) Deployment

4 Training Environment

The primary training environment was equipped with a GIGABYTE AORUS GeForce RTX 4090 MASTER 24G graphics card, featuring NVIDIA's state-of-the-art RTX architecture and advanced WINDFORCE cooling technology. This configuration was selected to guarantee optimal stability and efficiency throughout the training process. The training of the large model was completed over a duration of 101 hours, encompassing a total of 104,000 steps. The training for Hakka language phonetics took 23,000 steps and required 23 hours and 51 minutes.

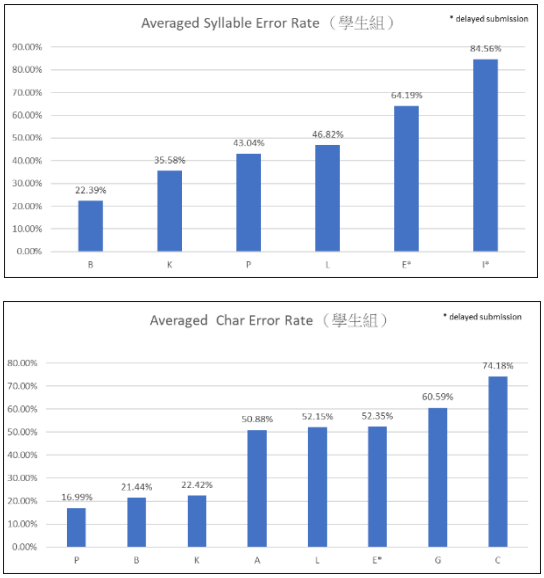
5 Results

The study successfully developed an Automatic Speech Recognition (ASR) model for

the Hakka language, leveraging the capabilities of Open AI's Whisper technology. When trained with online Hakka speech resources, the model demonstrated impressive proficiency. Model Performance Metrics (Team B):

Syllable Error Rate (SER): 22.39%

Character Error Rate (CER): 21.44%



Data Source: Formosa Speech Recognition Challenge 2023 - Hakka ASR[1]

In the competition, our performance metrics displayed noteworthy accomplishments. Specifically, in the Track 2 student group, our model achieved the top rank (1st place) for the Hakka pinyin syllable error rate. Meanwhile, in the Track 1 student group, we secured the second rank (2nd place) for the Chinese character error rate.

6 Conclusion

Developing a Hakka speech recognition model that is proficient in identifying Hakka nuances is a crucial step in promoting and conserving the rich Hakka language and heritage. By integrating this model into daily engagement with the Hakka language, a renewed technological zest will be introduced, enhancing its utility and allure. This study will integrate this model into various scenarios and platforms, expanding its reach and significance. Hakka cultural assets can be preserved and nurtured through language technology. The rapid pace of technological innovation and globalization makes it increasingly necessary to protect and rejuvenate minority languages, including Hakka. A state-of-the-art Hakka language automatic speech recognition system was developed by integrating data from various online repositories and OpenAI's Whisper technology. Ultimately, this system will lay the foundation for developing intelligent living solutions in Hakka pedagogy. The compelling results, as well as our accolades in the competition, speak volumes about the robustness and potential of the model. Thus, this study transcends mere scholarly pursuits; it offers a practical tool that can revolutionize Hakka-centric digital education. By integrating it into advanced applications, it is expected to significantly enhance the preservation and appreciation of Hakka traditions in the digital age. The recognition we garnered in competitive platforms solidifies our confidence in the model's resilience and its potential for shaping the future of language technology and cultural preservation.

Acknowledgments

This research is supported by the National Science and Technology Council, Taiwan, R.O.C. under Grant No. NSTC 112-2410-H-239 -015 -MY2. This study also thanks the organizers of the Formosa Speech Recognition Challenge 2023 and the Hakka Affairs Council for providing invaluable data for this study. Regarding the development of speech recognition technology, we would like to thank Dr. Hung-Shin Lee (North Co., Ltd.) for his professional consultation. For the collection and understanding of Hakka corpus, we extend our gratitude for the professional assistance provided by the following individuals: Li-Fen Huang (黃麗芬, Hakka Language Teacher at Taipei First Girls High School) and Fon-Siin Qi (徐煥昇, Teacher at Yu Ying Elementary School, Long Teng Branch).

References

Gandhi, S. (2022). Fine-Tune Whisper For Multilingual ASR with Transformers. Retrieved from https://huggingface.co/blog/fine-tune-whisper

[1] Formosa Speech Recognition Challenge 2023 - Hakka ASR, https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2023

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Paper presented at the International Conference on Machine Learning.

Scharenborg, O., Ciannella, F., Palaskar, S., Black, A., Metze, F., Ondel, L., & Hasegawa-Johnson, M. (2017). *Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: Preliminary results.* Paper presented at the Proc. Internat. Conference on Natural Language, Signal and Speech Processing (ICNLSSP).

Shah, P., Chadha, H. S., Gupta, A., Dhuriya, A., Chhimwal, N., Gaur, R., & Raghavan, V. (2022). Is Word Error Rate a good evaluation metric for Speech Recognition in Indic Languages? *arXiv preprint arXiv:2203.16601*.